DOCUMENT RESUME

ED 147 892 CS 501 916

TITLE Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, July 1 - December 31, 1977).
INSTITUTION Haskins Labs., New Haven, Conn.
REPORT NO SR-51/52-1977
PUB DATE Dec 77
NOTE 238p.

EDRS PRICE MF-$0.83 HC-$12.71 Plus Postage.
DESCRIPTORS *Acoustic Phonetics; *Articulation (Speech); *Artificial Speech; Audition (Physiology); *Auditory Perception; Language Handicaps; *Language Research; Lateral Dominance; Music; Phonemes; Phonetics; *Psychoacoustics; Reading Difficulty

ABSTRACT

This report is one of a regular series about the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The 17 papers discuss the identification of sine-wave analogues of speech sounds; prosodic information for vowel identity; progressive changes in articulatory patterns in verbal apraxia; temporal coordination of phonation and articulation in a case of verbal apraxia; factors in the maintenance and cessation of voicing; the influence of tempo on stop closure duration as a cue for voicing and place; reading reversals and developmental dyslexia; noncategorical perception of tone categories in Thai; effect of speaking rate on vowel formant movements; effects of transition length on identification and discrimination along a place continuum; perceptual integration and differentiation of spectral information across intervocalic stop closure intervals; musical skill and categorical perception of harmonic mode; phonetic and auditory aspects of adaptation; hemispheric specialization in children with language deficiency; source/tract coupling and vowel differences; vowel and consonant identification when cued by the same acoustic information; and information conveyed by vowels. (AA)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED147892

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SR-51/52 (1977)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 July - 31 December 1977

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports).

CS507 772

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development
Grant HD-01994

Assistant Chief Medical Director for Research and Development,
Research Center for Prosthetics, Veterans Administration
Contract V101(134)P-342

United States Army, Department of Defense
Contract MDA 904-77-C-0157

National Institutes of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
Biomedical Research Support Grant RR-5596

National Science Foundation
Grant BNS76-82023

National Science Foundation
Grant MCS76-81034

National Institute of Neurological and Communicative
Disorders and Stroke
Grant NS13870

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Thomas Baer
Fredericka Bell-Berti
Gloria J. Borden*
Guy Carden*
Robert Crowder*
Steven B. Davis
Michael Dorman*
Donna Erickson
William Ewan*
Carol A. Fowler*
Frances J. Freeman*
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Alice Healy*
David Isenberg
Leonard Katz*
Isabelle Y. Liberman*
Leigh Lisker*
Ignatius G. Mattingly*
Seiji Niimi[1]
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Donald P. Shankweiler*
Michael Studdert-Kennedy*
Michael T. Turvey*
Robert Verbrugge*
Hirohide Yoshioka[1]

Technical and Support Staff

Eric L. Andreasson
Elizabeth P. Clark
Sunila V. Dandekar
Donald Hailey
Terry Halwes
Elly Knight*
Sabina D. Koroluk
Agnes McKeon
Nancy R. O'Brien
Robin Rowedder*
William P. Scully
Richard S. Sharkany
Leonard Szubowicz
Edward R. Wiley
David Zeichner

Students*

William Balch
Steve Braddon
David Dechovitz
Laurel Dent
Susan Lea Donald
F. William Fischer
Hollis Fitch
Anne Fowler
Carole E. Gelfer
Janette Henderson
Nieba Jones
Lynn Kerr
Morey J. Kitzman
Andrea G. Levitt
Roland Mandler
Leonard Mark
Nancy McGarr
Georgia Nigro
Mary Jo Osberger
Abigail Reilly
Helen Simon
Emily Tobey
Betty Tuller
Harold Tzeutschler
Michele Werfelman

---

*Part-time
[1]Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

On the Identification of Sine-Wave Analogues of Certain Speech Sounds -- Peter J. Bailey, Quentin Summerfield and Michael Dorman . . 1

Prosodic Information for Vowel Identity -- Robert R. Verbrugge and Donald Shankweiler 27

Progressive Changes in Articulatory Patterns in Verbal Apraxia: A Longitudinal Case Study -- Elaine Sands, Frances J. Freeman and Katherine S. Harris 37

Temporal Coordination of Phonation and Articulation in a Case of Verbal Apraxia: A Voice Onset Time Study -- Frances J. Freeman, Elaine S. Sands and Katherine S. Harris 47

Factors in the Maintenance and Cessation of Voicing -- Leigh Lisker 55

The Influence of Tempo on Stop Closure Duration as a Cue for Voicing and Place -- Robert F. Port 59

Reading Reversals and Developmental Dyslexia: A Further Study -- F. William Fischer, Isabelle Y. Liberman and Donald Shankweiler . . . 75

The Noncategorical Perception of Tone Categories in Thai --Arthur S. Abramson 91

Effect of Speaking Rate on Vowel Formant Movements -- Thomas Gay . . 101

Effects of Transition Length on Identification and Discrimination Along a Place Continuum -- David Dechovitz and Roland Mandler 119

Perceptual Integration and Differentiation of Spectral Information Across Intervocalic Stop Closure Intervals -- Bruno H. Repp 131

Musical Skill and the Categorical Perception of Harmonic Mode -- Mark J. Blechner 139

Phonetic and Auditory Aspects of Adaptation: Evidence from Thai Voicing Contrasts -- S. Lea Donald 175

Hemispheric Specialization for Speech Perception in Kindergarten Children with Language Deficiency -- Davida R. Rosenblum and Michael F. Dorman 183

Can the Intrinsic F_0 Differences Between Vowels Be Explained by Source/Tract Coupling -- William G. Ewan 197

On the Relationship Between Vowel and Consonant Identification When Cued by the Same Acoustic Information -- Paul Mermelstein 201

Information Conveyed by Vowels: A Confirmation -- David Dechovitz 213

II. Publications and Reports 223

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-49) 225

I. MANUSCRIPTS AND EXTENDED REPORTS

On the Identification of Sine-Wave Analogues of Certain Speech Sounds*

Peter J. Bailey[†], Quentin Summerfield[††] and Michael Dorman[†††]

ABSTRACT

There is no obvious psychoacoustic basis for the perceptual categorization of synthetic stop consonant-vowel syllables according to place of production. Using a task that did not involve overt labeling, we compared the categorization of series of speech sounds formed by varying the onsets of the second and third formant transitions with the categorization of series of analogues of these sounds in which the formants were replaced with frequency- and amplitude-modulated sine-waves. These sine-wave patterns are perceived either as complex tones or as speechlike sounds in which a whistle is initiated by a stop consonant. Different category boundary positions accompanied these two percepts. When the sine-wave series were heard as speechlike, category boundaries were similar to those obtained with the formant stimuli. The demonstration that the perceptual categorization of an acoustic pattern is not determined solely by its spectro-temporal specification is discussed in the context of theoretical accounts of the distinction between speech and nonspeech.

INTRODUCTION

A concern of research that seeks to examine the abilities of human infants and nonhuman animals to perceive speech sounds, is that phonetic categories be dissociated from other auditory categories as the basis for the

---

*A partial summary of these results was presented at the 93rd Meeting of the Acoustical Society of America, State College, Pennsylvania, June 1977.

[†]Also Department of Psychology, The University of York, U.K.

[††]Also The M.R.C. Institute of Hearing Research, The University of Nottingham, U.K.

[†††]Also Speech and Hearing Clinic, Arizona State University at Tempe.

Acknowledgment: This work was carried out while Peter Bailey and Quentin Summerfield were supported by N.A.T.O. Post-doctoral Research Fellowships from the U.K. Science Research Council. We should like to thank Rod McGuire for his time and programming expertise, Erika Hamm for running the subjects in Experiment II, Michael Studdert-Kennedy for his comments on an earlier draft of this manuscript and, in particular, Alvin Liberman for his hospitality, advice and encouragement.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

measured response. Several studies have demonstrated that human neonates categorically perceive voiced and voiceless initial stop consonants (for example, Eimas, Siqueland, Jusczyk and Vigorito, 1971; Lasky, Syrdal-Lasky and Klein, 1975; Streeter, 1976). The interpretation that this ability reflects sensitivity to phonetic categories has been challenged on the grounds that the positions of phonetic and auditory category boundaries were confounded in the stimuli used (Stevens and Klatt, 1974; Kuhl and Miller, 1975; Miller, Wier, Pastore, Kelly and Dooling, 1976; Streeter, 1976; Pisoni, 1977). For example, the members of a continuum formed by varying the relative onset times of two coterminous tones are perceived in three categories with category boundaries corresponding to differences in onset time of about 20 msec (Pisoni, 1977). With tone-onset-times (TOTs) of less than about 20 msec, the onsets of the two tones are perceived as simultaneous; for TOTs of more than 20 msec, the onsets are perceived as successive. These auditory categories of simultaneity and successivity could underly the infant's ability to discriminate synthetic exemplars of voiced and voiceless stop consonants. However, while there are good grounds for contesting the claim that infants discriminate voicing contrasts phonetically, the situation with contrasts of place of production appears to be different. Both Eimas (1974) and Miller and Morse (1976) have demonstrated with similar stimuli, but different methodologies, that infants can discriminate place of production contrasts in initial position categorically, an ability for which a psychoacoustic rationale is less obvious. The stimuli for these experiments were three-formant consonant-vowel syllables generated by a parallel resonance synthesizer and are schematized in Figure 1.

English-speaking adults identify patterns A and B as [gæ] and patterns C, D and E as [dæ] (Pisoni, 1971). Eimas (1974) showed with nonnutritive sucking as a measure of habituation-dishabituation that infants discriminated patterns B and C but not D and E. Miller and Morse (1976) used a heart-rate measure of habituation-dishabituation in a within-subjects design to show that infants discriminate patterns B and C but not patterns A and B or C and D. These authors have noted that the differences in onset frequencies of the second and third formants are the same in the within-category pairs as in the between-category pairs. The results might be explained by arbitrarily according a special status to spectrally diverging patterns (for example, A and B); alternatively, one might contrive the hypothesis that of the five patterns only A and B exceed some threshold for spectral change in the infant's auditory system, although we are not aware of any empirical support for the latter suggestion. Perhaps the most obvious categories for auditory patterns of this kind are those corresponding either to no frequency change or to frequency change in a particular direction; however, the data of Eimas (1974) and of Miller and Morse (1976) suggest that infants do not use such categories. Thus, on the basis of this evidence and in the absence of evidence to the contrary,[1] the tentative conclusion that infants discriminate place of production contrasts phonetically appears to be justified. Nevertheless, we felt that the importance of accurately detailing the ontogeny of human perceptual abilities requires that alternative explanations be

---

[1]Popper (1972) claimed that discontinuities in the presumed auditory representation of the members of three-formant vowel-consonant place continua underly the location of category boundaries. This claim appears not to be supported by his own data.

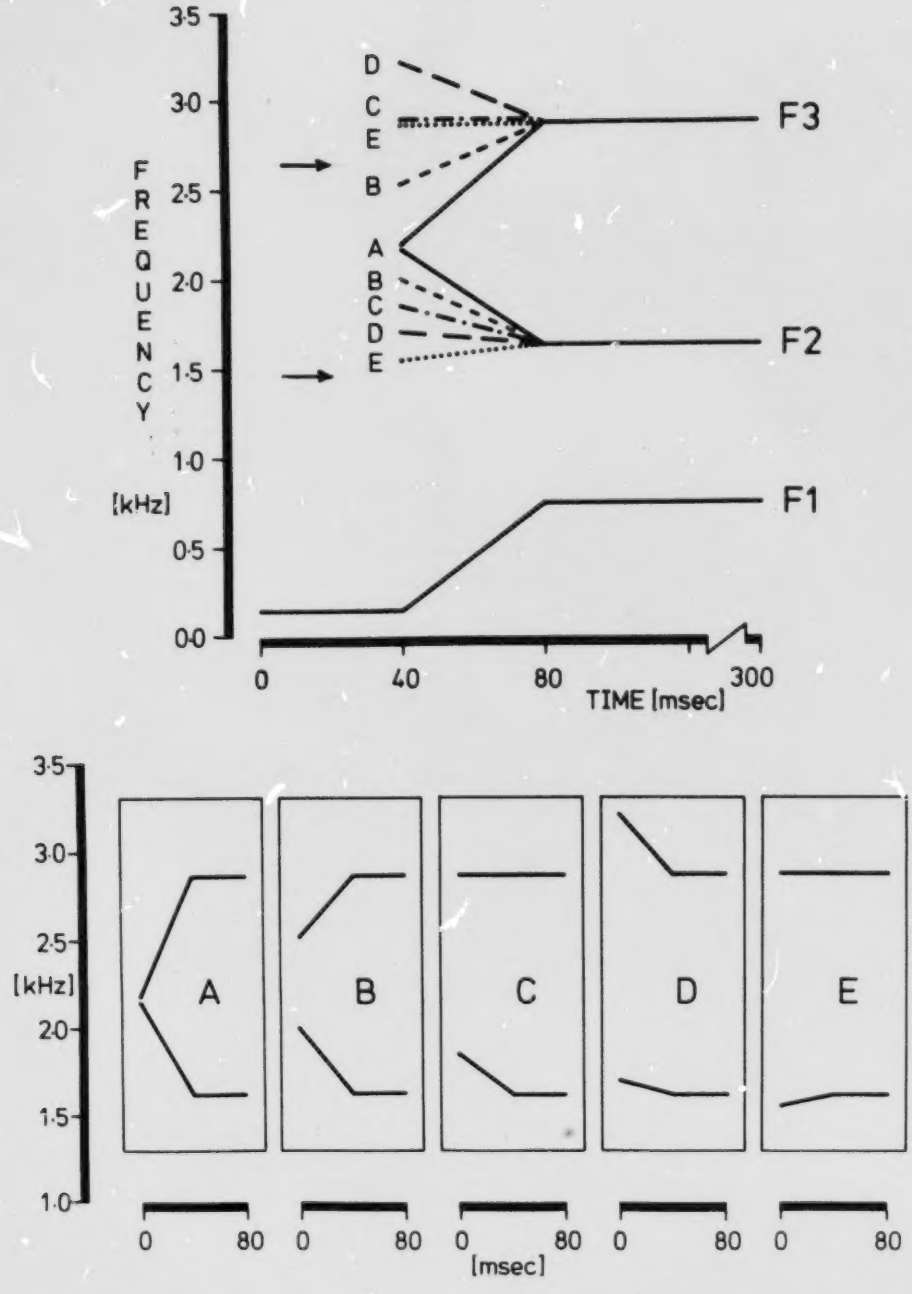

Figure 1: Schematic representations of the five stimulus patterns from Pisoni (1971) used by Eimas (1974) [B,C,D and E] and Miller and Morse (1976) [A,B,C,D]. See the text for details. The lower panel shows the second and third formant transitions in the five patterns individually. For clarity the transitions are represented here as linear segments while in the actual stimuli they were parabolic.

3

actively eliminated. As a contribution to this end, we wanted to determine whether any psychoacoustic rationale might be found to explain the categorization of three-formant speech patterns into bilabial and alveolar categories.

There are a number of ways in which this problem might be approached. Techniques in which specific psychophysical questions are asked using nonhuman animals (for example, Sinnott, Beecher, Moody and Stebbins, 1976) or infants as subjects possess the virtue of directness, but can usefully be supplemented by experimentation with human adults whose communicative abilities allow more detailed questions to be asked. We chose the latter course.

EXPERIMENT I

The object of Experiment I was to determine whether the categorization of acoustic patterns resembling consonant-vowel syllables is a function only of their spectro-temporal specification or whether they are heard as speechlike or as nonspeech.

METHOD

Stimuli

We synthesized two consonant-vowel (CV) continua with the OVE IIIc serial resonance synthesizer at the Haskins Laboratories.[2] Their spectro-temporal specifications are detailed in Figure 2. The continua were composed of three-formant patterns. (The synthesizer produces five formants. The two highest formants were removed by filtering as described below.) The parametric specifications of the first and third formants were the same in each continuum; only the second formant differed between the two. The configuration in the left-hand panel of Figure 2, with the steady state of the second formant (F_2) at 950 Hz, produces a vowel like [o]. The configuration in the right-hand panel, with the steady state of F_2 at 1800 Hz, produces a vowel like [e]. With each vowel a nine-member [b] to [d] continuum was created by covarying the onset frequencies of the second and third formants symmetrically about their steady states in eight 50 Hz steps. The third formant had its steady state at 2500 Hz, and the third formant transition ranged in onset from 2300 Hz to 2700 Hz. In the [bo-do] continuum, the onset of the F_2 transition ranged from 750 Hz to 1150 Hz. In the [be-de] continuum, the onset of the F_2 transition ranged from 1600 Hz to 2000 Hz. The first formant had its onset at 200 Hz and its steady state at 500 Hz. All formant transitions were 35 msec in duration with linear trajectories. The total duration of each syllable was 250 msec.

These parameter values were selected so that both continua would be physically symmetrical but phonetically asymmetrical. This is possible because the relative invariance of the point of constriction in the vocal tract for a particular place of production across vocalic contexts gives rise to a relatively invariant pattern of resonances corresponding to the closed vocal tract for a given place. These "locus frequencies" of approximately 800 Hz

---

[2]The synthesizer had been modified to ensure that the first pitch pulse occurred at the same point in every syllable.

FIGURE 2

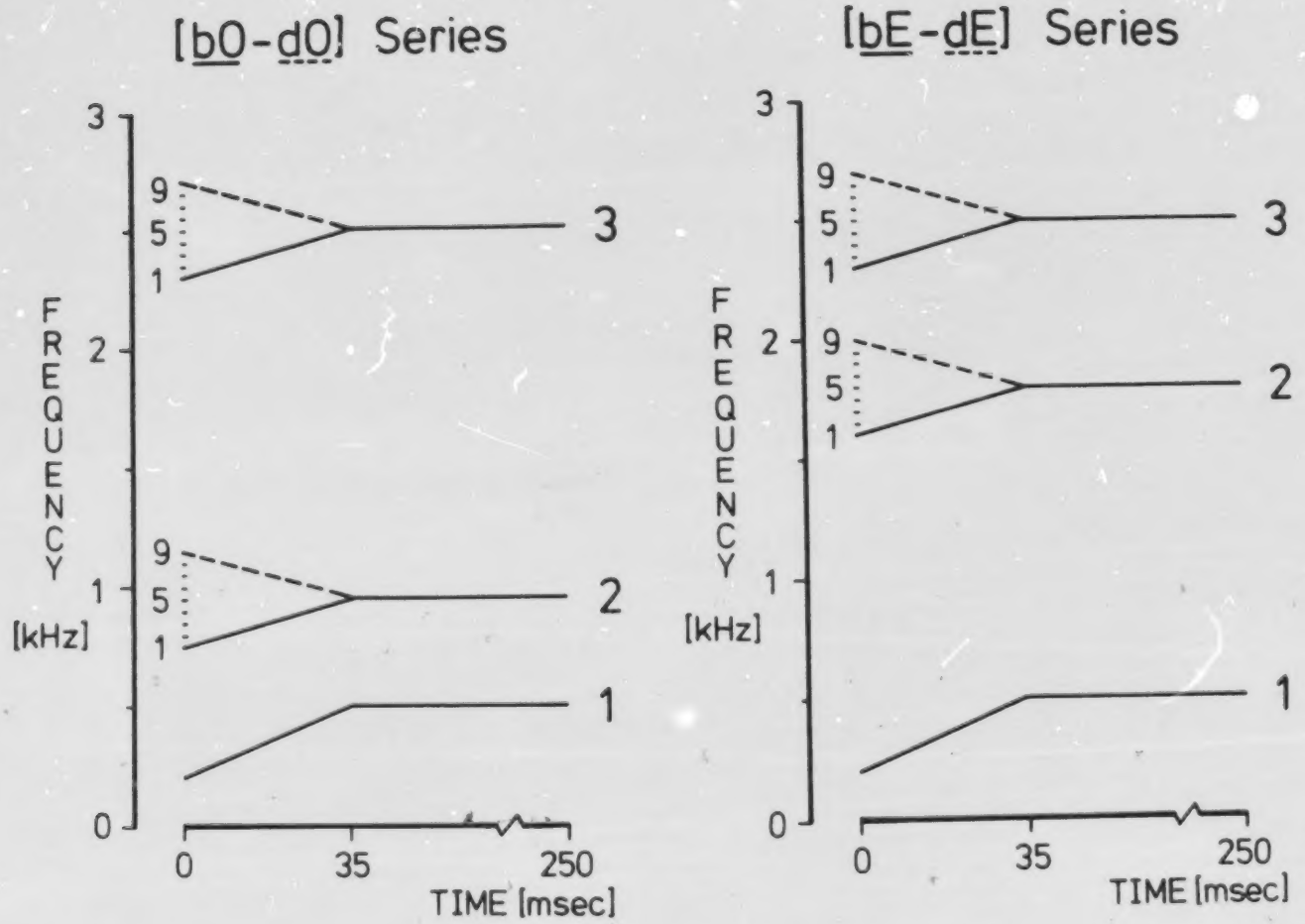

Figure 2: Schematic representations of the stimuli used in Experiment I. See the text for details.

for [b] and 1800 Hz for [d] (Delattre, Liberman and Cooper, 1955) are not realized in the signal. The initial formant transitions represent that portion of the change from the closure frequencies to those corresponding to the steady-state vowel during which there is sufficient airflow to excite the vocal tract. The relative frequencies of locus and steady-state determine the direction of formant transitions characterizing a stop with a particular place of production in the context of the following vowel. We intended the phoneme boundary on the [bo-do] continuum to be associated with falling transitions so that the majority of its members would be heard as [bo]. The inverse should apply to the [be-de] continuum; the majority of its members should be heard as [de]. In this way we intended to dissociate phonetic boundaries from any boundaries that might either accompany flat, as opposed to rising or falling, transitions or simply coincide with the centers of stimulus ranges.

These eighteen stimuli were low-pass filtered with a cut-off at 3.2 kHz and digitized with a sampling rate of 10 kHz. This operation eliminated the fourth and fifth formants. A hardware spectrum analysis was performed on successive 12.8 msec segments of each stimulus over the duration of its formant transitions, and the relative levels of the three formants were measured in each segment. Using these values to specify amplitudes, and using the parameters from the formant synthesis to specify frequencies, we copied the members of both continua by replacing their formants with frequency- and amplitude-modulated sine-waves. These sine-wave patterns were created with a digital software synthesizer and were recorded and redigitized in the same way as the formant stimuli. On first acquaintance these patterns sounded like nonspeech whistles.

Thus, the stimuli for Experiment I formed four continua: two were of vowel-type [o] corresponding to the patterns in the left-hand panel of Figure 2, and two were of vowel-type [e] corresponding to the patterns in the right-hand panel of Figure 2. For each vowel-type, one continuum was constructed from formants and one from sine-waves.

We wished to discover whether psychoacoustic boundaries determined with the sine-wave stimuli would coincide with phonetic boundaries determined with the formant stimuli. This required a task that can provide identification data without involving overt labeling. We used an AXB identification task. On each trial a triad of stimuli was presented. The first and third members of the triad were the end-point stimuli of a continuum, the second member of the triad was a stimulus drawn randomly from that continuum. The task of a listener was to indicate whether this random stimulus was more like the first or more like the last member of the triad. In addition, we wanted to obtain a conventional two-step AXB discrimination function for each continuum. In this test A and B were separated by two steps along the stimulus continuum and X was identical to either A or B.

For the identification test we recorded two sequences of trials for each continuum. The first was a practice sequence of three blocks of eighteen trials involving only the continuum end-points. Feedback was given on each trial in the form of the word 'first' or 'last' included on the tape as appropriate after the response interval. The second sequence consisted of nine blocks of eighteen trials of which the first three were intended to be practice blocks to be discarded before data analysis. In each block all members of the continuum were represented twice in the central position of the

AXB triad, once when A was stimulus #1 and B was stimulus #9, and once with the order reversed. On every trial the interval between the members of the triad was 1 sec, and an interval of 3 secs intervened between successive triads. An extra three seconds separated successive blocks of triads. The same timing pattern was incorporated in the AXB discrimination sequences for which we recorded a sequence of 6 blocks of 14 practice trials with feedback and a sequence of 9 blocks of 14 trials without feedback. Again, the first three of these blocks were intended for practice. For any continuum, each of the seven two-step pairs appeared equally often in each of its four possible orderings.

Subjects

Six undergraduates were paid to take part in the experiment. They declared themselves to be phonetically naive, to have normal hearing in both ears and to have learned English as their first language in the U.S.A. They were tested in a quiet room, either singly or in pairs in four two-hour sessions that were held on different days. All stimuli were presented binaurally through headphones at a level of 75 dB.

Procedure

In the first session subjects only listened to the sine-wave stimuli, which, they were informed, differed at their onsets. They described these sounds as nonspeech whistles. They were told that each trial of the experiment would consist of three stimuli and that their task was to decide whether the center stimulus sounded more like the first or more like the last member of the triad. They made their responses by writing down one of the letters 'F' or 'L' on a specially prepared response sheet. In this first session, subjects listened twice to both identification and discrimination practice sequences with feedback. The order of vowel-type was counterbalanced. Initially, subjects found the task difficult, but their performance improved during the session to at least 75 percent correct on the identification task.

In the second session the subjects again only listened to the sine-waves. They performed the discrimination test for each vowel-type and then the identification test for each vowel-type. Each test was preceded by a practice sequence with feedback. Four of the subjects heard the [e] analogues before the [o] analogues, while only two heard the [o] analogues before the [e] analogues. (The experiment had originally been designed for eight subjects, but only six were tested.)

In the third session subjects heard the formant stimuli in the same format, except that they listened to the discrimination tests for a second time at the end of the session. (Having examined the data from session 2, we realized that the number of discriminations in a single administration of the discrimination tests was insufficient to yield interpretable results.)

In the fourth session subjects once again only heard the sine-waves, but at the begining of this session the relationship between the formant and the sine-wave stimuli was explained. After listening to the sine-waves again, all subjects agreed that these stimuli could be heard as initiated by one of the stop consonants [b] or [d]. (In fact, one listener had made this observation

without prompting near the end of session 2.) Hearing the stimuli in this new way, subjects found the tasks easier, and during the session reported no tendency for the consonantal percept to disappear.

Results

The data were sorted to yield identification functions and discrimination functions for each subject in each condition. The results obtained in session 3 with the formant stimuli are shown in Figure 3. The graphs in the upper panel correspond to the identification test, those in the lower panel to the discrimination test. The data of all 6 subjects have been pooled to obtain these graphs and each point plots the mean of 72 responses in the identification test and of 144 responses in the discrimination test. In both graphs the solid line corresponds to the functions obtained with the [bo-do] continuum, and the dotted line corresponds to those obtained with the [be-de] continuum. For each continuum, the identification function relates the percentage of times each stimulus was judged to be more like the [b] end-point (that is, stimulus #1) than the [d] end-point (that is, stimulus #9) of its continuum. The stimulus numbers are arrayed along the horizontal axis, the percentage of [b]-like identifications increases along the vertical axis. Our expectations of phonetic asymmetry are at least partially borne out: the two functions do not overlap in the boundary region. The phoneme boundary on the [bo-do] continuum is displaced to the right of the center of the stimulus range; similarly, though to a very much lesser degree, the boundary on the [be-de] continuum is displaced to the left. The discrimination functions do not show any major peaks, but there is a tendency for the [be-de] stimuli to be discriminated better at lower stimulus numbers and for the [bo-do] stimuli to be discriminated better at higher stimulus numbers.

The identification and discrimination functions of Figure 3 can be compared with those for the sine-wave continua from session 2 that are displayed in Figure 4; the parameters of this display are the same as those of Figure 3, except that the data from the one subject who began to hear the stimuli as speechlike in session 2 have been excluded. Thus, each point in the identification function and in the discrimination function plots the mean of 60 responses. Again, two functions of each type have been plotted: the solid line corresponds to the analogues of the [bo-do] continuum, and the dotted line corresponds to the analogues of the [be-de] continuum. The pattern of data in this figure is very different from that displayed in Figure 3. The two sine-wave identification functions largely overlap throughout their ranges, and the 50 percent points on both functions fall close to the centers of the continua. No clear pattern emerges from the discrimination functions which are more variable than those obtained from the formant stimuli; as noted above, only half as many discrimination data per point were collected with the sine-wave stimuli.

Normal-ogive psychometric functions were fitted to the identification data of each subject in each condition using probit analysis (Finney, 1971). An estimate of the position of the phoneme boundary was obtained by computing the 50 percent point on each fitted function. For the formant stimuli, the mean of the boundaries on the [bo-do] continuum was 5.57, and on the [be-de] continuum it was 4.54. The difference between these means, although in the predicted direction, is not significant when assessed in a one-way analysis of variance ($F_{1,5}=3.13; p < 0.1$). Nevertheless, it is larger than the difference

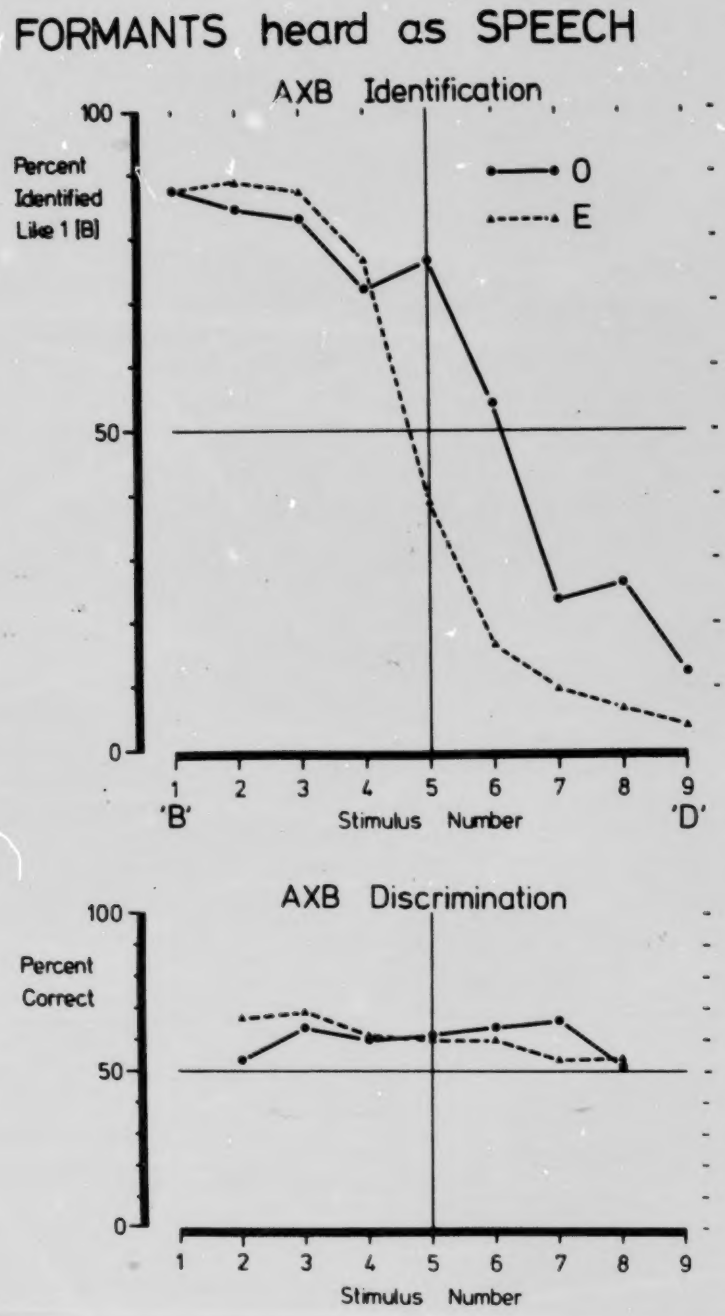

Figure 3: Identification and discrimination functions from Session 3 in Experiment I.

between the means obtained with the sine-wave stimuli that were (for the five subjects whose data are illustrated in Figure 4) 4.47 for the [bo-do] analogues and 4.42 for the [be-de] analogues. The difference between these means is not significant ($F_{1,4}$=0.01;p > 0.2).

The failure of the two formant continua to produce significant differences in mean boundaries is surprising. We can trace this outcome to two factors: the members of the [bo-do] continuum were not identified very systematically, and there was little asymmetry in the way the [be-de] continuum was identified. These factors reflect our failure to estimate stimulus parameters appropriately for naive listeners. However, it is possible to carry out a less stringent test of the asymmetry hypothesis by determining whether the members of the [bo-do] continuum were perceived as more [b]-like than the members of the [be-de] continuum. Using the number of [b]-like responses made to each stimulus by each subject as the dependent measure, an analysis of variance with the factors Subjects[6] x Vowel-types[2] x Stimuli[9] was performed. The effect of vowel-type was significant ($F_{1,5}$=12.31;p < 0.025), showing that overall, a greater number of [b]-like responses were made to the members of the [bo-do] continuum.

A similar analysis was carried out on the sine-wave data provided by the five subjects who heard the sine-waves as nonspeech sounds, in which the effect of vowel-type was not significant ($F_{1,4}$=0.69;p > 0.1). This is consistent with the pattern of data in Figure 4 and confirms the tendency for sine-wave continua to be divided symmetrically when heard as nonspeech sounds. This result is not materially influenced by the fact that our stimulus continua are only physically symmetrical when frequency is represented on a linear scale. Asymmetries introduced during peripheral auditory transmission, by the critical band mechanism (Scharf, 1970) for instance, would be likely to enhance the discriminability of rising, as oppposed to falling transitions, a distinction that is not apparent in our data. What is not revealed is the reason why subjects divide the sine-wave continua in the way they do. There are at least three possibilities: listeners may detect the presence or absence of spectral change in the higher resonances; they may distinguish rising from falling transitions; or they may simply divide the continua into two approximately equal ranges. However, none of these strategies can account for the asymmetrical categorization of the formant continua. This could either be correlated with the different spectral properties of formants and sine-waves, or with the way these stimuli are heard. The condition run in the fourth session of the experiment in which the sine-waves were heard as speechlike goes some way to dissociating the two possibilities.

The identification and discrimination data obtained in the fourth session are displayed in Figure 5. The pattern of these data is very different from that in Figure 4. Here, where the sine-waves were heard as speechlike, the identification and discrimination functions corresponding to the [o] and [e] analogues no longer overlap. Moreover, as with the formant stimuli, the majority of the [bo-do] analogues were heard as [b]-like, while the majority of the [be-de] analogues were heard as [d]-like. In this condition, the discrimination functions were reasonably smooth and consistent and were similar to those obtained with the formant stimuli. As before, boundaries were estimated by probit analysis. The mean boundary obtained with the [bo-do] analogues corresponded to a stimulus number of 5.65. That obtained with the [be-de] analogues was 3.82. The difference between these means is

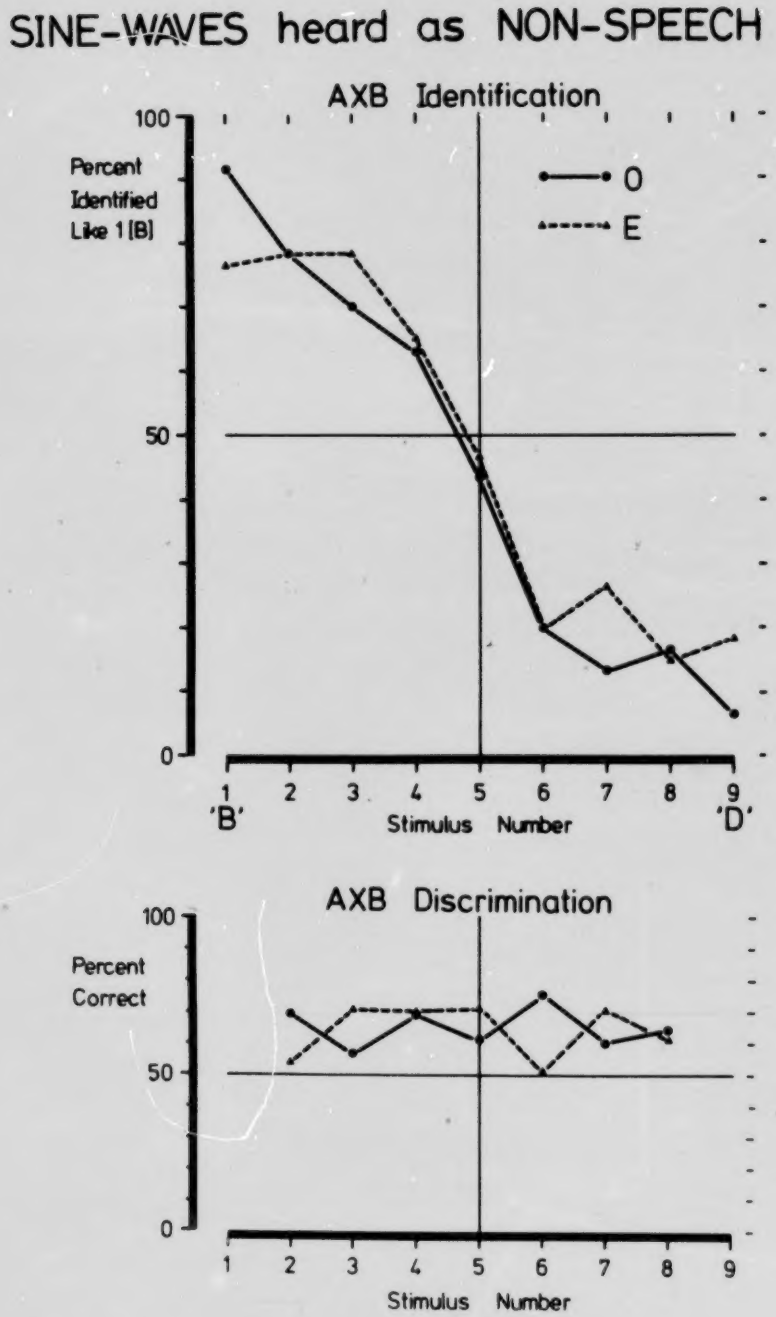

Figure 4: Identification and discrimination functions from Session 2 in Experiment I.

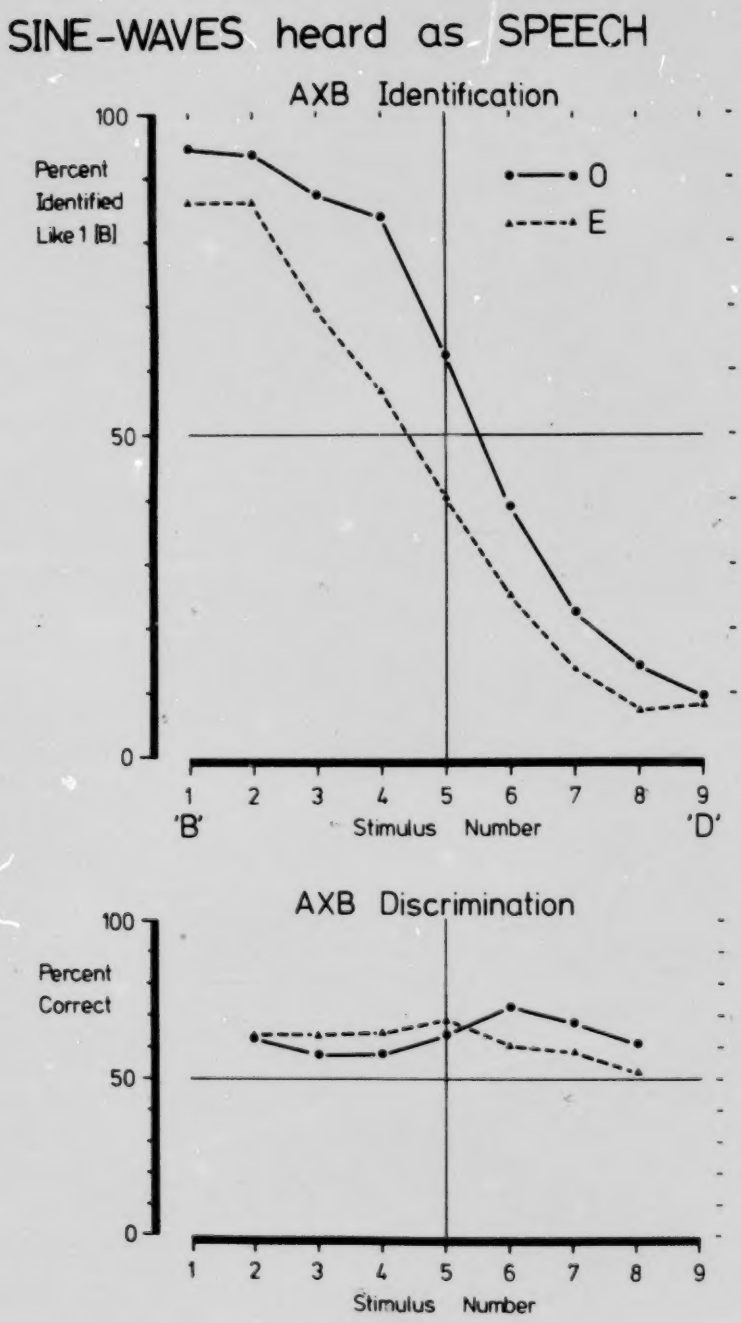

Figure 5: Identification and discrimination functions from Session 4 in Experiment I.

significant ($F_{1,5}$=18.35;$p < 0.001$). The pattern of data obtained in session 4 when the sine-waves were heard as speechlike is clearly more akin to the pattern corresponding to the formant stimuli than to that obtained in session 2 when the sine-waves were heard as nonspeech whistles. This correspondence suggests that the results are due not so much to the spectral structure of the stimuli, but rather to the way in which they are heard. Formants and sine-waves produce similar patterns of data when both are heard as speech.

We should be explicit about what we mean when we say that the sine-waves could be heard as 'speechlike'. We do not mean that after repeated exposure to these sounds and to their formant analogues listeners were able to identify their onset frequencies and infer a correspondence to [b] or [d]. Rather, it is our own experience and that described by our listeners, that the sine-wave patterns can 'name themselves' (cf. Studdert-Kennedy, 1976, p. 244). One hears an initial [b] or [d] followed by a whistle. We say that the percept is 'speech-like' then because, while the stop is compellingly phonetic, what follows it is not. In order to determine to what extent this result was idiosyncratic of these six listeners or was a function of exposure to formant stimuli in Session 3, we ran a second experiment.

EXPERIMENT II

For this experiment we synthesized a single [ba-da] continuum. Again we copied these formant stimuli with frequency- and amplitude-modulated sine-waves. We intended the new continuum to be both more natural and more asymmetrical than the continua used in Experiment I. As a result, more listeners heard the sine-wave analogues as speechlike without prompting or prior exposure to the formant stimuli. Thus, we were able to divide our subjects into two groups on the basis of their descriptions of their initial perception of the sine-wave stimuli.

METHOD

Stimuli

A single 11-member [ba-da] continuum was created with the OVE IIIc synthesizer. The total duration of each stimulus was 250 msec, and the duration of the syllable-initial formant transitions was 40 msec. The first formant had its onset at 350 Hz and rose linearly to a steady state at 750 Hz. The second and third formants had their steady states at 1000 Hz and 2500 Hz, respectively. The onset of the F_2 transition ranged from 650 Hz to 1350 Hz in ten 70 Hz steps; the onset of the F_3 transition ranged from 2250 Hz to 2750 Hz in ten 50 Hz steps. Thus, as in Experiment I, the frequency transitions in this continuum ranged symmetrically about the steady states. The stimuli were low-pass filtered at 3.2 kHz and digitized at a sampling rate of 10 kHz. The waveforms were analyzed with a hardware spectrum analyzer, and the relative levels of the formants measured over the durations of the formant transitions. As before, these measurements were used to control the levels of three frequency-modulated sine-waves produced by a digital synthesizer. In this way a sine-wave analogue of each member of the [ba-da] continuum was created.

We recorded two AXB identification tests with 110 trials in each, one with formant stimuli and one with sine-wave stimuli. The format of these

tests was the same as that used in Experiment I, except that only 0.5 sec intervened between successive members of each triad. No discrimination tests were administered in this experiment.

Subjects and Procedure

Thirty subjects were tested. They were drawn from the members of a course in speech and hearing at Arizona State University. The AXB task was explained to the subjects and they were told that they would hear stimuli constructed from sine-waves, but they were told nothing about the relation between the sine-waves and possible speech sounds. The test with the sine-waves was administered first, followed by the test with formant stimuli. At the end of each test, subjects were instructed to write down a description of the stimuli.

Results

Somewhat fortuitously, 15 listeners said that they heard the sine-waves as nonspeech whistles or tones, while 15 listeners heard them as speechlike. The latter group described the sine-waves as being initiated by either a stop consonant or semi-vowel with bilabial or alveolar place of production.[3]

Figure 6 displays the AXB identification data for the group who said that they heard the sine-waves as nonspeech sounds. The dotted line plots the function obtained with the sine-wave stimuli; the solid line plots the function obtained with the formant stimuli. The formant continuum was divided asymmetrically into two sharply segregated categories. The sine-wave continuum, on the other hand, was divided less asymmetrically and into two less sharply defined categories. The contrast between the way the continua were categorized is exemplified by the responses to stimulus #6, the stimulus with flat transitions in the second and third resonances. When represented by formants, this stimulus was identified as like the [b]-endpoint of its continuum on 98 percent of its presentations; when represented by sine-waves, it was identified in this way on only 56 percent of its presentations. However, it is clear that the sine-wave continuum was not divided into two equal ranges, in contrast to the results of Experiment I. Figure 7 displays the identification functions for the group who said they heard the sine-waves as speechlike. These subjects also divided the formant continuum asymmetrically. Figure 7 suggests that sine-waves heard as speechlike are categorized in a similar way to formant stimuli. This similarity is exemplified by a comparison of responses to stimulus #6 in each continuum: when represented by formants, this stimulus was identified as like the [b] end-point of its continuum on 88 percent of its presentations; when represented by sine-waves, it was identified this way on 85 percent of its presentations.

We examined the statistical reliability of these observations in several ways. First, two measures were extracted from psychometric functions fitted to the identification data of each subject in each condition. One was the

---

[3]Of the fifteen subjects who heard the sine-wave stimuli as speech sounds, five perceived [b] and [d], six [w] and [d], two [w] and [l], one [w] and [z] and one [w] and [y].

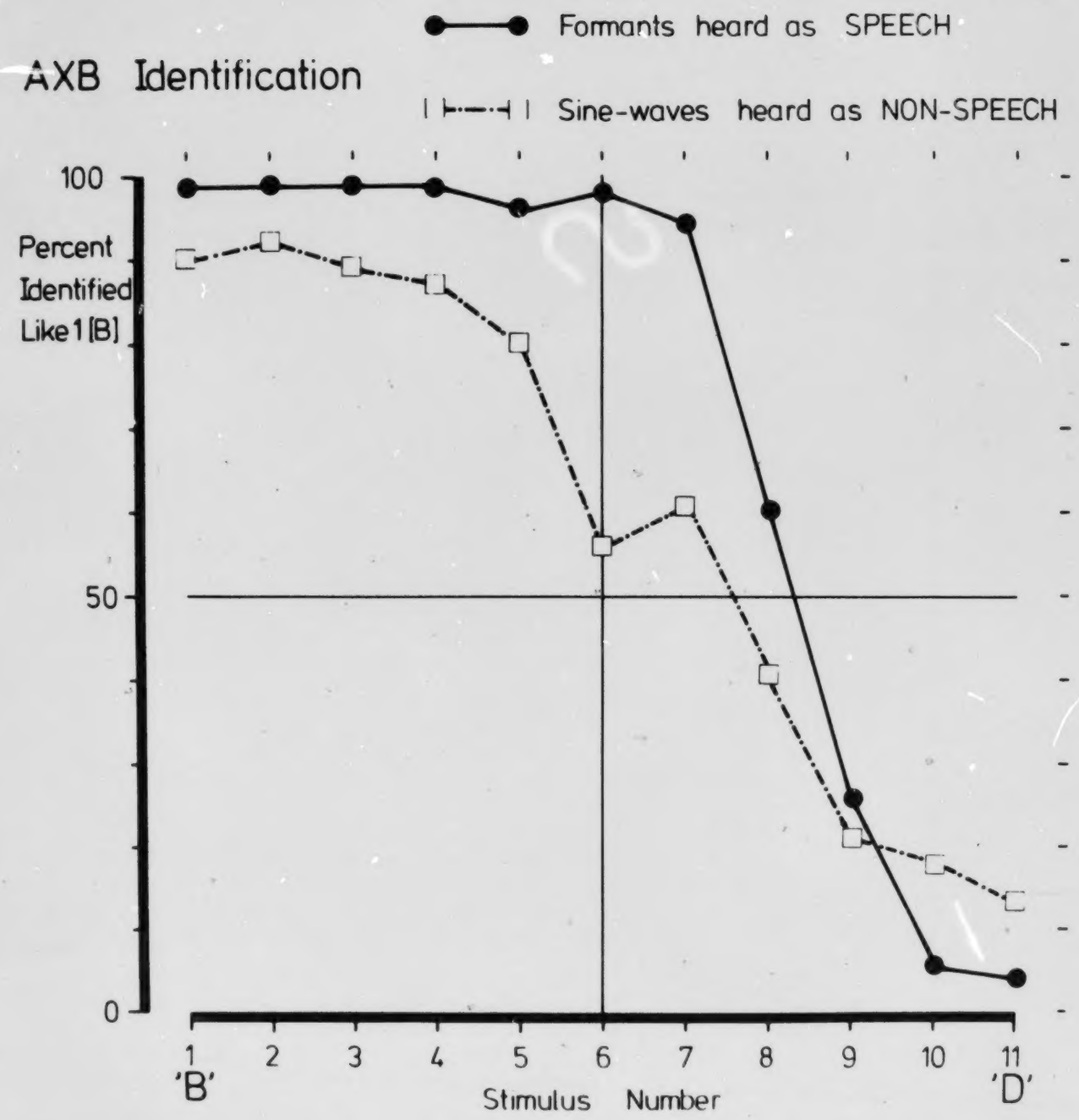

Figure 6: Identification functions for the group who heard sine-waves as speech-like in Experiment II.

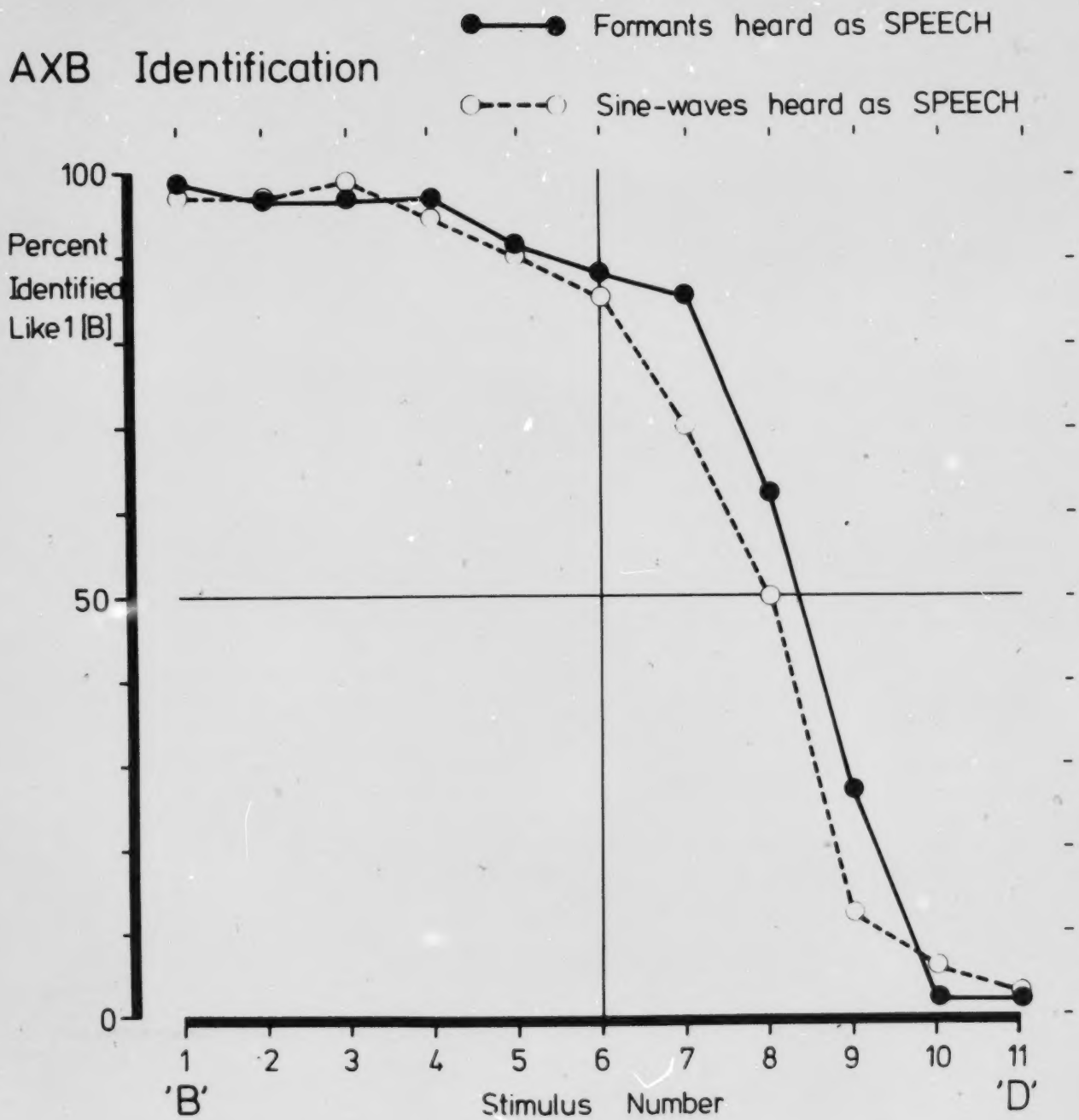

Figure 7: Identification functions for the group who heard sine-waves as non-speech in Experiment II.

16

stimulus number corresponding to the 50 percent point on the fitted function: this provides an estimate of the position of the phoneme boundary. The other was the slope of the probit regression line, a parameter that is directly related to the slope of the identification function in the boundary region. For the group who heard the sine-waves as nonspeech, category boundaries corresponded to stimulus numbers of 7.01 for the sine-wave stimuli and 8.25 for the formant stimuli. For the group who heard the sine-waves as speech-like, the equivalent values were 7.57 and 7.93. (The central stimulus in each continuum was number 6.) The differences between these means were assessed in an analysis of variance that showed that the overall difference between the boundaries on the two continua is significant ($F_{1,28}=12.74; p < 0.01$); however, the size of this difference does not differ significantly between the two groups of subjects ($F_{1,28}=3.79; 0.1 > p > 0.05$).

A similar analysis was undertaken with probit regression line slopes as the dependent measure, which showed that identification functions produced by sine-wave stimuli were flatter than those produced by formant stimuli ($F_{1,28}=13.27; p < 0.01$). However, in this case, the size of the effect does differ between the two groups ($F_{1,28}=6.44; p < 0.025$). To ensure that this interaction does not result from differences between the slopes of the formant functions, we carried out analyses comparing the two groups for the formants and sine-waves independently. Only the sine-wave slopes differ between the groups ($F_{1,28}=4.24; p < 0.05$), confirming that flatter slopes were produced when sine-waves were heard as nonspeech.

Despite the failure of the first of these dependent measures to distinguish the condition in which the sine-waves were heard as nonspeech from the other conditions, this distinction is implied by the patterns of data in Figures 6 and 7. In an attempt to specify how the two groups of subjects differ, we carried out two further analyses, using as the dependent measure the number of [b]-like responses made by each subject to each stimulus. For the group who heard the sine-waves as speechlike, the mean percentages of [b]-like responses made to the formant and sine-wave continua were 69.34 percent and 64.00 percent (difference = 5.34 percent). For the group who heard the sine-waves as nonspeech sounds, the means were 70.91 percent and 58.97 percent (difference = 11.94 percent). In an analysis with the factors Groups (2)xContinua(2), the interaction between these two factors was significant ($F_{1,28}=4.59; p < 0.05$), showing that the difference between the formant and the sine-wave means is significantly larger for the group who heard the sine-waves as nonspeech. However, we cannot conclude from this that significantly fewer [b]-like responses were made to the sine-wave stimuli, because the groups also differed in the number of [b]-like responses that they made to the formant stimuli. Therefore, we ran analyses to compare the means for the sine-waves and formants independently. Although the two formant means do not differ significantly ($F_{1,28}=1.32; p > 0.2$), neither do the two sine-wave means ($F_{1,28}=2.55; p > 0.1$).

Discussion

The results of Experiment II are consistent with one aspect of the results of Experiment I. In both experiments, the categorization of sine-wave stimuli depended upon how the sine-waves were heard and, when the sine-waves were heard as speechlike, more closely approximated the categorization of formant stimuli. Taken together, these results indicate that the way in which

a sound is categorized is not simply a function of the spectral structure of that sound, but also relates to how that sound is heard. However, there is one respect in which the results of the two experiments do differ. In Experiment I, the sine-wave continua were divided symmetrically when heard as nonspeech. In the equivalent condition of Experiment II, the continuum was not divided symmetrically.

It is possible that these different results follow from differences in the proximities of the first and second resonances in the stimuli used in the experiments. The first resonance was closer to the second resonance in the [ba-da] continua used in Experiment II than it was in either the [bo-do] or the [be-de] continua used in Experiment I. If the upward spread of masking from the first resonance has the effect that equal differences in the onset frequency of the second resonance become more discriminable as the onset frequency of the second resonance rises, then the result found when the sine-waves were heard as nonspeech in Experiment II might be expected. One reason for questioning this explanation, however, is that in the sine-wave continuum, as in the formant continuum used in Experiment II, the intensity of the second resonance at its onset frequency increased as its onset frequency was lowered. This might have been expected to counteract the effects of masking.

GENERAL DISCUSSION

Let us first consider the results obtained when the sine-wave continua were not heard as speechlike. The equivocal outcomes of our two experiments provide evidence both for and against an auditory discontinuity underlying the location of category boundaries in the perception of place of production. This unsatisfactory situation must be resolved by further experimentation. However, the demonstration of symmetrical categorization of sine-wave continua when heard as nonspeech in Experiment I, is consistent with the results of experiments on the discrimination of formant transitions in nonspeech contexts. Mattingly, Liberman, Syrdal and Halwes (1971) found no discrimination peaks corresponding to the location of phoneme boundaries when second formant transitions extracted from the members of a [bæ-dæ-gæ] continuum were presented in isolation. We have already noted that there is no obvious auditory strategy for dividing continua of frequency transitions that can account consistently for the way subjects categorize formant and sine-wave continua when these are heard as speech. We now turn to these data.

To a greater or lesser extent, category boundaries on formant and sine-wave continua, when heard as speechlike, were placed asymmetrically. In Experiment I the [be-de] sine-wave continuum and both formant continua were categorized asymmetrically by five listeners out of six, while the [bo-do] sine-wave continuum was categorized asymmetrically by four listeners out of six. The results of Experiment II, on the other hand, are unequivocal in showing asymmetrical categorization of the [ba-da] continuum. All thirty subjects categorized the formant continuum asymmetrically, while fourteen out of the fifteen subjects who heard the sine-wave continuum as speechlike categorized it asymmetrically.

The finding that phoneme boundaries are not tied to particular directions of formant movement is not restricted to our data (for example, Liberman, Delattre, Cooper and Gerstman, 1954; Delattre, et al., 1955; Pisoni, 1971; Blumstein, Stevens and Nigro, 1977). For instance, the two arrows in Figure 1

indicate the onset frequencies of the second and third formants at the mean [b-d] phoneme boundary found by Pisoni (1971). Here, for the vowel [æ], both transitions are rising at the boundary. This pattern can be contrasted with that found by Blumstein et al. (1977) for the vowel [a], where the boundary was characterized by a slightly rising transition in F_3 and a slightly falling transition in F_2. What explanation can account for the placement of category boundaries on continua of synthetic syllables that vary in place of production?

One explanation has emphasized the correspondence of general properties of the acoustic signal to discrete phonetic categories (Stevens, 1975). For instance, Blumstein et al. (1977) observe that: "Acoustic energy at the stimulus onset is spread or 'diffuse' for the [d] and [b] stimuli and is concentrated in a narrow frequency region or is 'compact' for the [g] stimuli. These properties may be described in terms of the onset characteristics and the following spectral changes: a diffuse-rising pattern characterizing [b], a diffuse-falling pattern for [d] and a compact-spreading pattern for [g]" (p. 1036). We find it reasonable to suppose that the abilities to produce and perceive speech have coevolved in such a way that maximally different acoustic patterns support the information for discrete categories of articulatory events. However, the data reported in the present paper and those reviewed above, are not compatible with the idea that there is a one-to-one isomorphism between the registration of these acoustic properties and the perception of phonemic identity. Whether or not this isomorphism exists in the acoustic description of natural productions, the need to account for the perception of synthetic speech remains. For example, we need to characterize the commonality which underlies a bilabial percept, whether this results from frequency-modulated sine-waves, synthetic formant transitions, or the rich acoustical structure of natural speech. Indeed, we feel that an account of phonetic perception should be based on a rationalization of sensitivity to that commonality, rather than on an enumeration of sensitivities to specific acoustical elements in the speech stream.

An integral component of this rationalization is the dissociation of phonetic from nonphonetic perception. The original focus of our interest in these experiments was the realization that naive listeners, who initially describe the sine-wave patterns as whistles, come to perceive them phonetically, and that the change seems to be irreversible.[4] What underlies this changing percept of an unchanging stimulus? This question could be answered in many ways. One answer is provided by a class of models in which the perceptual system detects the same array of sensory information but processes that information in different ways. Two examples of this type of solution are considered below. We shall contrast them with an alternative view in which no explicit distinction is drawn between different modes of processing; the changing percept results from a change in the organization of attention to information in the signal.

---

[4]In an experiment by Cutting (1974) using sine-wave stimuli somewhat like those used in the present experiment, listeners apparently did not report that the stimuli took on a phonetic character. The differences in stimuli and procedure between his experiment and ours render the difference in outcome difficult to interpret.

Most accounts of speech perception, confronted with the apparent uniqueness both of the speech signal and of its perceptual consequences, have dichotomized sounds into two classes: speech and nonspeech. Having construed the dichotomy, they have had to explain how the perceptual process achieves this classification. As noted above, we can identify two types of solutions to the problem: in one, the decision about speech-likeness is assumed to be explicit and directive; in the other, the decision is implicit and passive. According to the former, speech-likeness is supposedly marked by specific acoustical attributes that, if detected in an initial stage of auditory analysis, direct the signal to a special phonetic processor. For example, Stevens and House (1972), following House, Stevens, Sandel and Arnold (1962), remark that "The listener need not be set for speech prior to his hearing the signal; his prepared state is triggered by the presence of a signal that has appropriate acoustic properties" (p. 13). We note that this account could be buttressed to explain the change from hearing sine-waves as nonspeech to hearing them as speechlike if supplemented with a variable criterion for the acceptability of the evidence provided by the first stage. An initially conservative setting of the criterion could be relaxed to achieve intelligibility within the context specified by an experiment. However, although the data can be explained in this way, we can question the general approach on two grounds. First, attempts to identify acoustical trigger features empirically have not been successful (Haggard, 1971; Allen and Haggard, 1977). Second, and fundamentally, it is hard to see how the phonetic processor could have evolved without the omniscience of the initial classification stage. Given that, why should the phonetic processor have evolved at all? Similar objections can be raised to the suggestion made by Cutting (1974) that a high-level decision as to the status of the signal might censor the output of a phonetic processor when the signal is insufficiently speechlike.

The other type of solution to the problem of distinguishing speech from nonspeech sounds starts by suggesting that phonetic and generalized auditory analyses are accorded in parallel to all acoustic inputs. Phenomenal perception corresponds to whichever process achieves a satisfactory analysis. For instance, Liberman, Mattingly and Turvey (1972) have suggested that "...the incoming signal goes indiscriminately to speech and nonspeech processors. If the speech processors succeed in extracting phonetic features, then the signal is speech; if they fail, then the signal is processed only as nonspeech" (pp. 323-324). Clearly this approach depends upon a characterization of the acoustical representation of phonetic features. Our demonstration of the perceptual duality of sine-wave stimuli can be accommodated in this model by a provision for context-sensitive adjustment in the specification of adequate stimuli for detectors of phonetic features. But, given such provision, we wonder whether the speech processor would ever abandon the search for phonetic features to admit a nonspeech solution.

We suggest that both of the preceding accounts of the distinction between speech and nonspeech are either inadequate or incomplete because they fail to capture an important inherent characteristic of the speech signal. In particular, they do not achieve explanatory adequacy because they assume that the information in the signal that must underpin phonetic perception is a specification only of discrete acoustic elements in the three-dimensional metric of frequency, amplitude and time, and not of the <u>origin</u> of the elements expressed in the four-dimensional metric of a three-dimensional vocal tract undergoing continuous reconfiguration over time. Those accounts which com-

mence with only a three-dimensional specification of the signal suppose that speech perception is mediated by knowledge of the way vocal tracts behave. For example, Stevens and House (1972) suggest that "After processing by peripheral auditory structures, some attributes of an incoming auditory pattern are then, as it were, looked up in the dictionary of auditory-articulatory correspondences" (p. 54); a similar well-developed view presented in detail elsewhere (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967) is illustrated by Mattingly and Liberman (1969): "In effect, the key that the listener has available to him is an articulatory model that relates the phonetic message to the signal" (p. 102). Seen in this way, perception consists of interpreting elements by imposing structure upon them, but note that this structure derives from constraints embodied in an internal articulatory model. A perceptual model of this kind would seem to involve at least two stages: in the first, a sequence of acoustic elements must be segregated and detected; in the second, these elements must be interpreted, presumably to reconstruct the information encoded in the sequential properties of the signal. Knowledge of vocal tract behavior may assist the first stage but it governs the second stage. While we have no doubt that speech perception is inextricably tied to the origin of the signal in a vocal tract, we wonder whether a process of fractionation followed by reintegration would best capture the information endowed to the signal by the continuous articulatory flow of a dynamic vocal tract.

In the following, we attempt to sketch in very general terms how an alternative account of the distinction between speech and nonspeech sounds might be developed. It will become clear that this account fits naturally into the wider context of a view of the perception of speech that might loosely be described as ecological. We shall discuss the general view while acknowledging that our data are only indirectly supportive of it.

In the natural world, sounds result from the participation of three-dimensional structures in events that occur over time. We suppose that the evolution of sensitivity to sound pressure variation progressed by a developing facility in identifying events that produce sounds, not just sounds per se (Gibson, 1966). When they specify events, acoustic signals describe not only their source but also what that source is doing. We do not yet have a sophisticated account of how this information is represented in the speech signal, but acoustic variation corresponding more or less directly to vocal tract cavity size variation can be identified, and perceptual sensitivity to it demonstrated (Kuhn, 1975). We wish to examine the possibility that the perception of acoustic patterns, whether speech or nonspeech, is properly described by registration of the coherence between information specifying the source of a sound and that specifying the transformation wrought upon the source. What we understand by coherence may be illustrated in a visual analogy. When a man runs, he structures light in such a way that both his identity as a man and his act of running are specified. When we perceive him running, we detect the coherence of these specifications; we do not first perceive the actor in order that we may interpret the elements of his act. Similarly we do not perceive speech by imposing articulatory structure upon an otherwise unstructured array of elements; rather, we perceive that structure because it is specified in the organization of the elements. This suggests an answer to the question of what is a speech sound: a pattern of sound may be perceived as speech if it specifies coherently its source as a human vocal tract partaking in a physiologically permissible act of articulation. The

registration of coherence is analogous to perceiving the solutions to a set of simultaneous equations. The equations provide structure and coherence for the solutions, but no one solution necessarily mediates the attainment of any other.

We require a more precise specification of what these notions entail, but we find them an appealing account of the way our listeners heard the sine-wave stimuli. When sine-waves were heard as nonspeech sounds, we suppose that listeners attended to the elements in the acoustic array but not to their organization. In hearing them as speechlike, on the other hand, they attended both to the elements and to their organization (cf. Polanyi, 1969), that together specify, albeit in a highly reduced form, a 'vocal tract' undergoing a bilabial or an alveolar articulation. Those familiar with R. C. James' photograph reproduced in Lindsay and Norman (1972, p. 8) will recognize that the foregoing analogously describes both the initial perception of the picture as a random array of dark and light areas and the subsequent perception of a Dalmatian dog walking in dappled sunlight. Both hearing sine-waves as speechlike and seeing the Dalmatian are compelling perceptions. Perhaps the search for coherence in stimulus information is a general goal of perceptual systems, guided and rewarded by the attainment of clarity (Gibson, 1966). We have already noted that when our listeners switched to hearing sine-waves as speechlike, their identification functions became more consistent and more categorical.

We have suggested that adult listeners perceive speech directly by obtaining information about articulation from an acoustic waveform. In the introduction to this paper, we reviewed experiments that show that infants and adults have similar sensitivities to place of production contrasts. Do infants, like adults, perceive articulatory events? The arguments above lead us to suppose that they do, and that the tendency to search for the coherence in sounds that specifies the events that produce them is an innate predisposition. For a human being, there would be considerable utility in a genetic endowment to detect the particular coherence found in the productions of human vocal tracts. Clearly, further data from three types of currently ongoing experimentation are required to evaluate these assertions. The first is the endeavor to specify how articulatory events structure sound in perceptually accessible ways. This may be achieved through examination of the correspondence between perceptual sensitivity and particular acoustic events in the speech signal (for example, Kuhn, 1975 and the present experiment), but a more fruitful way to specify the metric of the information in speech sounds may be to specify first the metric of articulatory dynamics (Fowler, 1977). The second type of experimentation seeks to delimit the sensitivity of the neonate to acoustic patterns having articulatory relevance (for example, Eimas, 1974; Miller and Morse, 1976) and to plot its ontogenetic refinement (for example, Simon, 1977). The third is experimentation with nonhuman animals reared in controlled environments that assesses perceptual sensitivity in circumstances where, we should predict, there is no innate predisposition to detect information about human articulation (for example, Kuhl and Miller, 1975; Sinnott et al., 1976). The evaluation of these data will be facilitated if attention is given to ensuring that the stimuli used in such experiments dissociate psychoacoustic and phonetic categories as the basis for the measured response.

SUMMARY

In the two experiments reported here, we attempted to determine whether a 'psychoacoustic' basis exists for the classification of continua of synthetic speech sounds that vary in place of articulation. What we mean by a psychoacoustic basis would be the existence of some attribute of the auditory system that predisposes the categorization of acoustic patterns into groups bearing a one-to-one correspondence with their phonetic labels. We studied the categorization both of continua of three formant CV syllables and of sounds modeled on these in which formants were replaced by frequency- and amplitude-modulated sine-waves. Our first experiment produced no support for the psychoacoustic explanation. The sine-wave continua were divided symmetrically into two halves with boundaries corresponding to flat initial transitions. The formant continua, on the other hand, were divided asymmetrically with boundaries corresponding either to rising or falling transitions. However, the results of the second experiment were equivocal: both formants and sine-waves were divided asymmetrically, although the formants were categorized more consistently and tended to be categorized more asymmetrically. Nevertheless, taken together, the results of the two experiments suggested that different information was detected when the sine-wave stimuli were heard as nonspeech and when the formant stimuli were heard as speech. This feeling was endorsed by the finding that the sine-wave stimuli could be perceived as speechlike: that is, they could be heard as initiated by a clear bilabial or alveolar consonant. When heard as speechlike, sine-waves were categorized more like formant stimuli, both more consistently and more asymmetrically than when they were heard as whistles. Thus the pattern of results appeared to relate not so much to the spectral structure of the stimuli, as to the way in which the stimuli were heard.

The perceptual duality of these sine-wave patterns provoked us to scrutinize existing accounts of the difference between speech and nonspeech. These accounts imply that speech is perceived when the acoustic elements in a sound stream can be interpreted by reference to an internalized representation of the vocal tract. We examined an alternative that supposes that the acoustic signal completely specifies the articulatory event that produces it. This account suggests that a sound is perceived as speech when it specifies coherently its source as a human vocal tract participating in a physiologically permissible act of articulation. We find this description of phonetic perception to be underspecified but appealing. From this orientation, the task both of the perceiver and of the experimenter is to determine how the acoustic signal specifies articulatory events.

REFERENCES

Allen, J. and M. P. Haggard. (1977) Perception of voicing and place features in whispered speech: a dichotic choice analysis. Percept. Psychophys. 21, 315-322.

Blumstein, S. E., K. N. Stevens and G. N. Nigro. (1977) Property detectors for bursts and transitions in speech perception. J. Acoust. Soc. Am. 61, 1301-1313.

Cutting, J. E. (1974) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.

Delattre, P. C., A. M. Liberman and F. S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.

Eimas, P. D. (1974) Auditory and linguistic processing of cues for place of articulation by infants. Percept. Psychophys. 16, 513-521.

Eimas, P. D., E. R. Siqueland, P. W. Jusczyk and J. M. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Finney, D. J. (1971) Probit Analysis. (Cambridge, U.K.: Cambridge University Press).

Fowler, C. A. (1977) Timing control in speech production. Ph.D. thesis, University of Connecticut.

Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).

Haggard, M. P. (1971) Encoding and the REA for speech signals. Quart. J. Exp. Psychol. 23, 34-45.

House, A. S., K. N. Stevens, T. T. Sandel and J. B. Arnold. (1962) On the learning of speech-like vocabularies. J. Verbal Learn. Verbal Behav. 1, 133-143.

Kuhl, P. A. and J. D. Miller. (1975) Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. Science 190, 69-72.

Kuhn, G. M. (1975) On the front cavity resonance and its possible role in speech perception. J. Acoust. Soc. Am. 58, 428-433.

Lasky, R. E., A. Syrdal-Lasky and R. E. Klein. (1975) VOT discrimination by four to six and a half month old infants from Spanish environments. J. Exp. Child Psychol. 20, 213-225.

Liberman, A. M., P. C. Delattre, F. S. Cooper and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of stop and nasal consonants. Psychol. Monogr. 68, no. 8 (whole no. 379).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., I. G. Mattingly and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Winston), pp. 307-334.

Lindsay, P. H. and D. A. Norman. (1972) Human Information Processing: An Introduction to Psychology. (New York: Academic Press).

Mattingly, I. G. and A. M. Liberman. (1969) The speech code and the physiology of language. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer), pp. 97-118.

Mattingly, I. G., A. M. Liberman, A. Syrdal and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Miller, C. L. and P. A. Morse. (1976) The "Heart" of categorical speech discrimination in young infants. J. Speech Hearing Res. 19, 578-589.

Miller, J. D., C. C. Wier, R. Pastore, W. J. Kelly and R. J. Dooling. (1976) Discrimination and labelling of noise-buzz sequences with varying noise-lead times: an example of categorical perception. J. Acoust. Soc. Am. 60, 410-417.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. (Ph. D. thesis, University of Michigan.) [Supplement to Haskins Laboratories Status Report on Speech Research].

Pisoni, D. B. (1977) Identification and discrimination of the relative onset times of two component tones: Implications for voicing perception in stops. J. Acoust. Soc. Am. 61, 1352-1361.

Polanyi, M. (1969) Knowing and being. In Knowing and Being, ed. by M. Greene. (Chicago: University of Chicago Press), pp. 123-137.

Popper, R. D. (1972) Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. J. Psycholing. Res. 1,

205-219.

Scharf, B. (1970) Critical bands. In Foundations of Modern Auditory Theory, ed. by J. V. Tobias, vol. 1. (New York: Academic Press), pp. 157-202.

Simon, C. (1977) Cross-language study of speech pattern learning. J. Acoust. Soc. Am. 61, S64(A).

Sinnott, J. M., M. D. Beecher, D. B. Moody and W. C. Stebbins. (1976) Speech sound discrimination by monkeys and humans. J. Acoust. Soc. Am. 60, 687-695.

Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and the Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (New York: Academic Press), pp. 303-330.

Stevens, K. N. and A. House. (1972) Speech perception. In Foundations of Modern Auditory Theory, ed. by J. V. Tobias, vol. 2. (New York: Academic Press), pp. 1-62.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.

Streeter, L. A. (1976) Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. Nature 259, 39-41.

Studdert-Kennedy, M. (1976) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press), pp. 243-294.

Prosodic Information for Vowel Identity*

Robert R. Verbrugge† and Donald Shankweiler†

ABSTRACT

Earlier research suggests that prosodic information is used by listeners in identifying vowels. The effects of variation in stress and tempo need to be distinguished. In this study, an adult male talker recorded nine American English monophthongs in /p-p/ syllables in each of four sentence contexts: two levels of syllable stress (stressed, destressed) were crossed with two rates of sentence articulation (slow, fast). Listeners made more errors when identifying syllables in isolation (average of 15 percent errors) than in sentence context (5 percent). Errors on the isolated syllables showed a bias toward "short" vowels, suggesting that the effective tempo perceived by listeners was slower than appropriate. Tokens of each syllable type were interchanged between sentence pairs so as to misinform listeners about the syllables' tempo and stress value. When tempo was mismatched, errors increased substantially on the fast syllables, showing a bias toward shorter vowels. No change was found when stressed and destressed syllables were interchanged, though formant frequency differences were greater than those between fast and slow syllables. The results contribute to a growing body of evidence that sentential stress-timing structure carries important information for vowel identity.

Most research on the perception of phonemes has focused on the information available in single syllables. Yet phonemes typically appear in prosodic patterns that extend over broader temporal ranges. It is important to consider whether the information specifying syllables in isolation is sufficient to define their identity in connected speech.

Our focus in the present study is the relationship between information for vowel identity and information for prosodic pattern. In earlier studies we have found that vowel identification is impaired when a destressed syllable

---

*This paper was presented at the 93rd meeting of the Acoustical Society of America, Pennsylvania State University, State College, Pennsylvania, 6-10 June 1977.

†Also University of Connecticut, Storrs.

Acknowledgment: The research was supported by a grant to Haskins Laboratories from the National Institute of Child Health and Human Development (NICHD Grant HD-01994; BRSG Grant RR-05596).

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

is excised from a sentence and heard in isolation (Verbrugge, Strange, Shankweiler, and Edman, 1976). Identification is also impaired when a citation-form syllable is spliced into the center of a full sentence. Thus, perception is most veridical when a vowel is heard in its original articulatory environment, whether an extended sentence or a single syllable. Our hunch has been that the original prosodic context specifies the rate of articulation, and that this allows listeners to compensate for the acoustic transformations produced by tempo variation. However, these studies may have confounded perceived tempo and perceived stress--listeners' errors might have been due to misinformation about syllable stress, rather than misinformation about rate. In the present study, we sought to compare the acoustic transformations produced by rate and stress variation, and to separate the effects of misinforming listeners about these two aspects of a syllable's prosodic identity.

An adult male talker recorded 36 sentences containing /p/-vowel-/p/ syllables. Nine American English monophthongs were spoken in each of four sentence types: the component /pVp/ syllable was either stressed or destressed, and the sentence was spoken at either a slow or fast rate of articulation. The sentence frame itself was lexically constant. For example, a stressed syllable appeared in the sentence "I think it's the yellow /piṕs/ call," while the target syllable was destressed in "I think it's the yellow /pips/ call."

The talker produced the two rate variants reliably: slow sentences averaged 1.9 seconds in duration, fast sentences averaged 1.3 seconds, and variability in durations was small. Not surprisingly, the durations of the fast /pVp/ syllables were consistently shorter than for the slow syllables, with duration measured from initial release to final release. However, the durations of the stressed and destressed syllables were roughly equal, suggesting that a sharp contrast in stress values had not been attained.[1]

Cross-sectional formant measurements told a complementary story. For each syllable we measured the center formant frequencies for the pitch periods showing maximum amplitude. Of course, these measures are only symptoms of the transformations produced by tempo and stress variation, but if these symptoms show divergent patterns in the two cases, it is likely that the transformations themselves are distinct. Figure 1 exhibits the spectral symptoms of stress variation. The nine vowels are plotted according to their first and second formant frequencies. Vowels in fast stressed syllables are identified by dashed lines, and fast destressed syllables by solid lines. The vowels show a consistent decrease in first formant values in the destressed syllables. Thus, the stress contrast was sufficient to produce a marked difference in formant values, even though overall syllable durations were equal.

---

[1]Note that the stress contrast investigated here involves the presence or absence of emphatic stress in the test syllables, not a contrast between stressed and unstressed syllables. The similarity of durations for the stressed and destressed syllables is probably symptomatic of a similarity in stress-timing patterns: the syllables occupied a full "metric foot" in both of the sentential stress patterns.

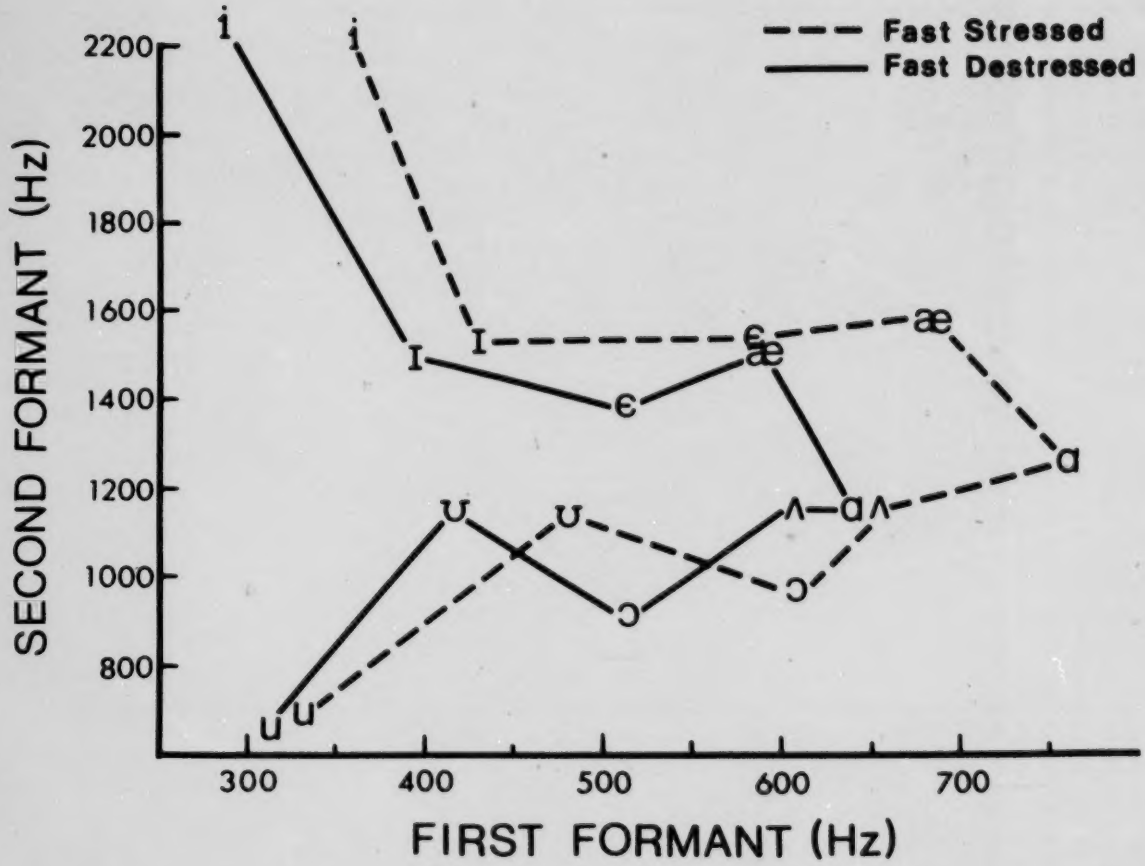

Figure 1: Center frequencies of first and second formants of nine vowels spoken in /pVps/ syllables: fast stressed and fast destressed conditions.

FIGURE 1

Figure 2 exhibits the spectral symptoms of tempo variation. Again the vowels from fast stressed syllables are identified by dashed lines. The solid lines now connect the measurements for slow stressed syllables. When speaking rate shifts from slow to fast, formant measures show neither the same pattern nor the same degree of change as was observed for destressing in Figure 1. Thus, in contrast to destressing, a rapid tempo did affect syllable durations, but had a comparatively small effect on cross-sectional formant measures. This difference in acoustic effects is compatible with the claim made by other investigators (for example, Gay, 1974; Gay, 1977) that the destressing and rate transformations are distinct.

To determine whether listeners attend differentially to these prosodic transforms when identifying vowels, we devised five listening tests. In one condition, listeners heard the syllables in their original sentence context; listeners heard several tokens of each sentence, in randomized order, and were asked to identify the /pVps/ syllable as one of nine alternatives. In a second condition, listeners heard the same syllables in isolation; the /pVps/ syllables were excised from the sentences by editing digitized versions of their waveforms. In the remaining three conditions, the excised syllables were transplanted into foreign sentence environments and these new sentences were assembled into listening tests. In the Rate Interchange condition, fast and slow syllables were interchanged between sentences, with the stress pattern held constant. In the Stress Interchange condition, stressed and destressed syllables were interchanged, with rate held constant. In the Both Interchange condition, the syllables were transplanted into environments in which both the rate and stress had originally been different.

Figure 3 presents the overall percent errors in identifying the syllables in these five conditions. Errors averaged 5 or 6 percent in the original sentence environment and in each of the three interchange conditions. The most dramatic contrast is between hearing the syllables in <u>any</u> sentence context and hearing them in isolation. Errors on the isolated syllables averaged 15 percent. It is clear from this contrast alone that sentential structure enhances the intelligibility of a component syllable, even when semantic and syntactic factors have been neutralized. The puzzle is why even an "incompatible" environment is better than none. One possible account is this: listeners may attend to the isolated syllables as citation-form syllables. If so, the effective mismatch between the event as perceived and as originally produced may be much greater than in any of the interchange conditions, resulting in a large increase in errors. This hypothesis is supported indirectly by the observation that slow stressed syllables showed comparatively few errors when heard in isolation; these syllables are acoustically similar to citation-form syllables, and thus can be perceived with little error as citation-form syllables.[2] The hypothesis is also supported by the types of errors listeners made. The errors in isolation did not reflect a global increase in confusions when compared to errors in sentence context; the increased errors were largely due to perceiving /pæps/ as /pɛps/, and /pɑps/ as /pʌps/. In each case listeners were strongly biased toward hearing the

---

[2]Errors on the slow stressed syllables averaged 6 percent when heard in isolation, compared to 15 percent for slow destressed syllables, 21 percent for fast stressed syllables, and 18 percent for fast destressed syllables.

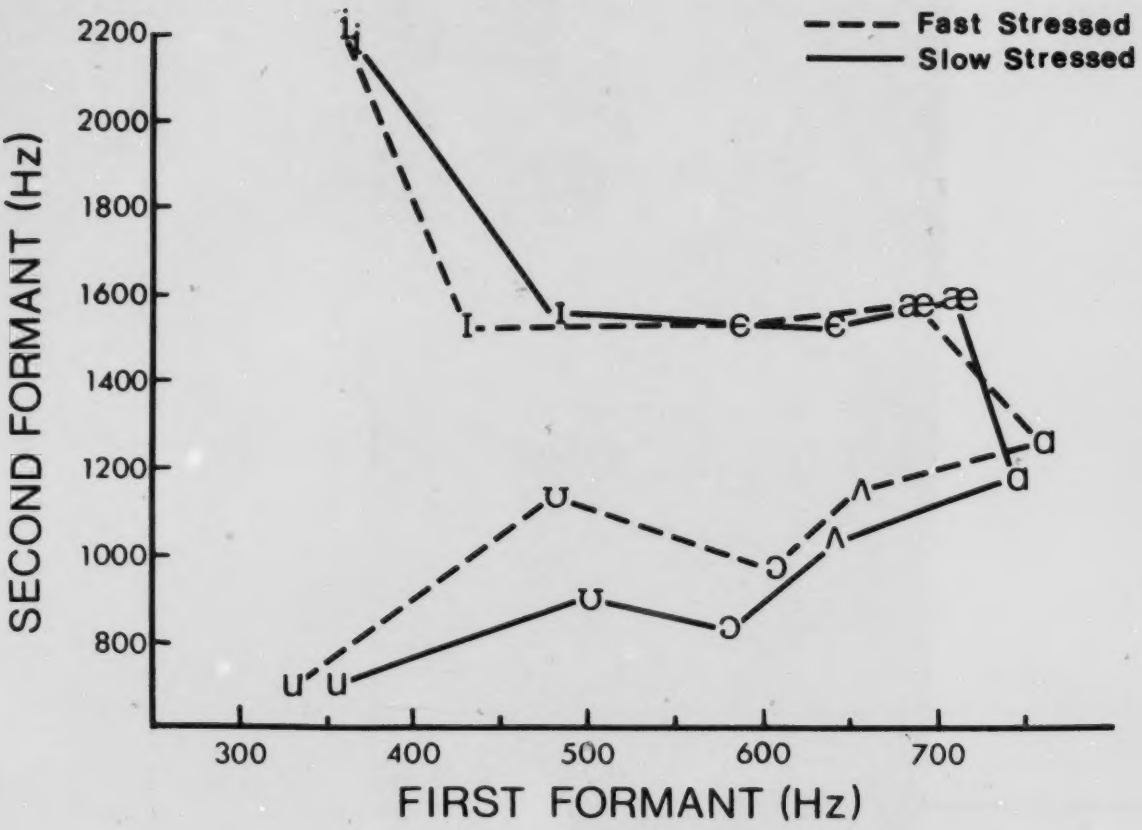

FIGURE 2

Figure 2: Center frequencies of first and second formants of nine vowels spoken in /pVps/ syllables: fast stressed and slow stressed conditions.

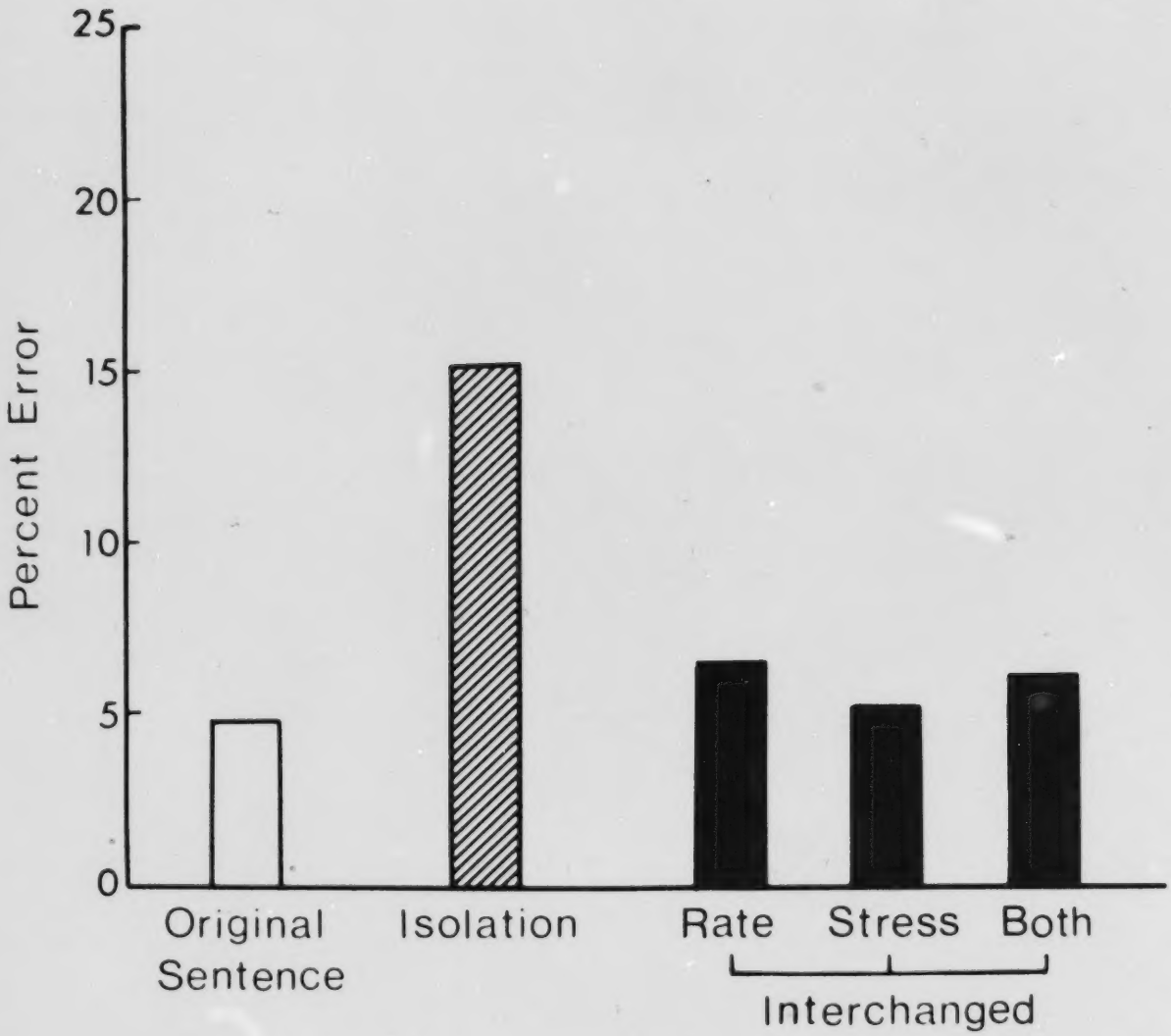

Figure 3: Mean percent errors in identifying vowels in /pVps/ syllables under five conditions: syllables heard in the original sentence, in isolation, in a sentence of different rate, a sentence of different stress value, and a sentence differing in both rate and stress values.

32

shorter vowel. This biasing was strongest for the fast excised syllables, suggesting that it was largely due to misinformation about the tempo of articulation when the syllables were presented as isolated events.

The averaged figures for the interchange conditions disguise some variation of interest. In Figure 4 the errors are divided between slow and fast syllables. Errors for each syllable type are depicted in open bars for the original sentence contexts, hatched bars for interchanged rate contexts, and solid bars for interchanged stress contexts. In the Rate Interchange condition there was a significant increase in errors for fast syllables. When the fast syllables were heard in a slow environment, there was a tendency for /pæps/ to be heard as /pɛps/, and /pɑps/ as /pʌps/. These errors were similar in kind to those found when the fast syllables were heard in isolation, again implicating temporal misinformation in the isolated syllable effect. There is an asymmetry in these rate effects that we can only note, without explanation: slow syllables do not show a complementary short-to-long-vowel bias in the rate interchange condition. For some reason they are more self-contained, less contextually sensitive, in specifying the component vowel.

In the Stress Interchange condition there were no significant shifts in listeners' identifications for any syllable types. It is worth noting that there were substantial formant frequency differences between stressed and destressed syllables, yet interchanging these syllables did not affect their perceived identity. It is clear that such spectral measures alone cannot predict the degree of confusion in cross-spliced conditions. The likely explanation for the failure to find a stress effect is that the perceived stress of the syllables was not successfully altered by their transfer to a new environment. The syllables themselves contain information for stress level and this apparently dominated listeners' perception of the overall stress pattern of the new sentences. There is both a methodological and theoretical lesson to be learned here: a simple transplant is not sufficient to alter a syllable's perceived stress value. The stress pattern of the sentence frame is not sufficient to override the constraints imposed by the syllable structure itself. Instead, what listeners perceive is a new stress pattern and a peculiar motivational state in the talker.

The fast and slow syllables present a mixed case from this perspective. Many of the fast syllables did not impose tight constraints on perceived rate of articulation and thus were more malleable to rate specification by their environment. On the other hand, the slow syllables appeared to be more tightly constrained, since they altered little in a new environment.

In general, it is clear that prosodic factors can exert an influence on vowel identification. In the study reported here, such effects were found only for speaking rate. Naturally this does not preclude finding stress pattern effects under different conditions. However, it does suggest that the effects we observed in earlier studies were due to tempo misinformation rather than to the stress values confounded with it.

How listeners' phonemic judgments are influenced by tempo remains an open question. Most approaches to the question assume that a speed or rhythm can be extracted from a target syllable's environment. This rate could then be used, for example, to scale the duration of the syllable, for comparison with

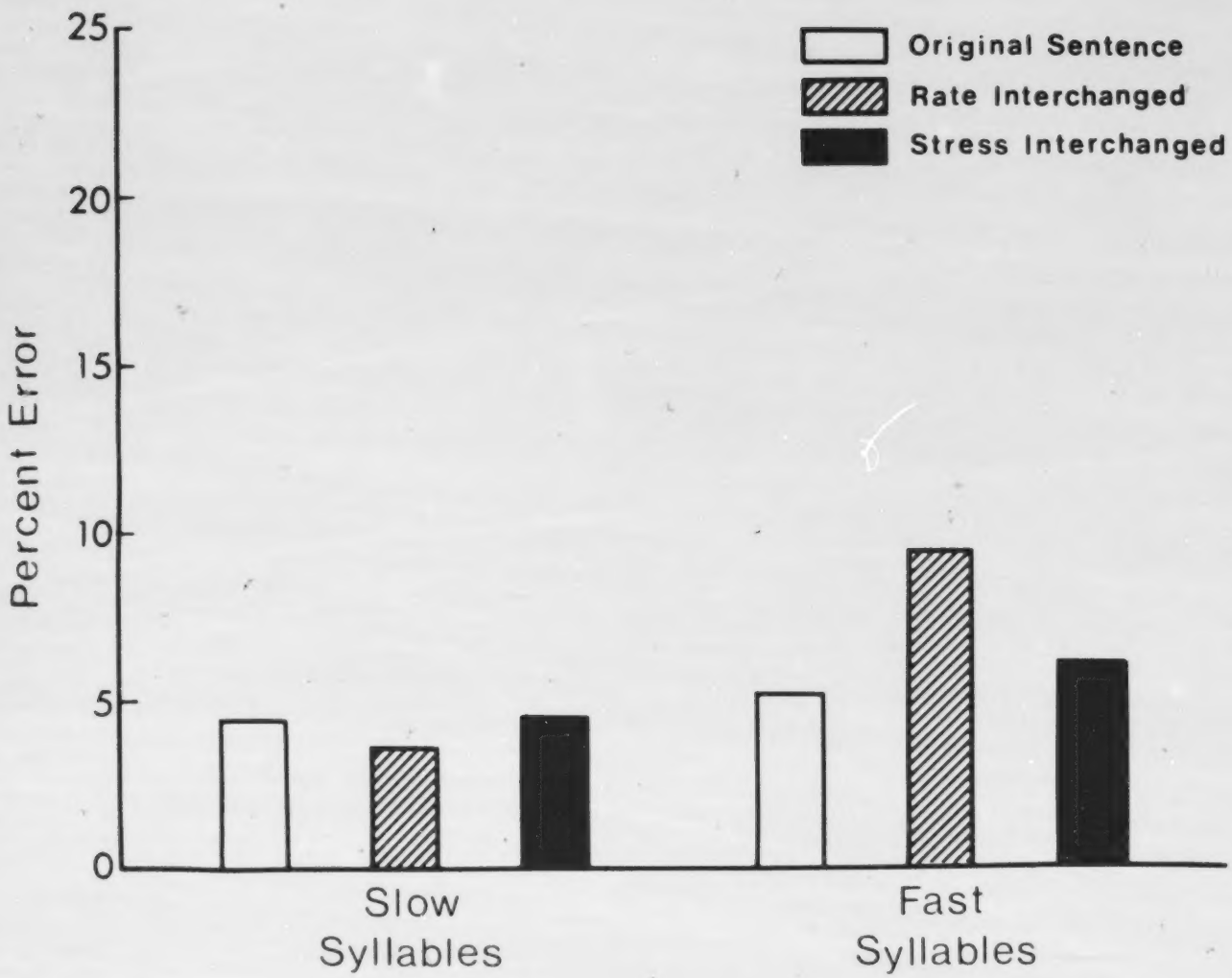

Figure 4: Mean percent errors in identifying vowels in slow and fast syllables in three sentential environments.

normative values of "intrinsic duration." Or the preceding context could permit some kind of entrainment to a speech rhythm, thereby facilitating segmentation and detection. These approaches tend to assume that some syllables (namely the "context") are determinate in both temporal and phonemic identity, and that other syllables are equivocal and needy of interpretation. This distinction seems gratuitous. These approaches also assume a stability of tempo and rhythm that is probably approached only in formal speech and metrical verse. The challenge for future research is to determine how the syllables in connected speech can cospecify both phonemic and prosodic structure, not only when rhythms are regular but also when tempo is continuously variable.

REFERENCES

Gay, T. (1974) A cinefluorographic study of vowel production. J. Phonet. 2, 255-266.

Gay, T. (1977) Effect of speaking rate on vowel formant movements. Haskins Laboratories Status Report on Speech Research SR-51/52.

Verbrugge, R. R., W. Strange, D. P. Shankweiler and T. R. Edman. (1976) What information enables a listener to map a talker's vowel space? J. Acoust. Soc. Am. 60, 198-212.

Progressive Changes in Articulatory Patterns in Verbal Apraxia: A Longitudinal Case Study*

Elaine Sands,† Frances J. Freeman†† and Katherine S. Harris†††

ABSTRACT

This research reports findings in a ten-year study of an apraxic adult, who was one of five subjects described by Shankweiler and Harris (1966). Confusion matrices and feature analysis were used to compare 1965 with 1975 performance. Results indicate that over the ten-year period, errors of place, manner and omission were markedly reduced. However, voicing errors, while reduced in total number, still constituted a significant percentage of the patient's residual errors. Implications of these findings are discussed.

INTRODUCTION

The nature of the articulatory disturbance that often accompanies aphasia has long been the subject of interest and disagreement in the literature. Disparate nomenclature (Aten, Darley, Deal and Johns, 1975; Martin, 1974), is but one example of the lack of consensus. The syndrome has been termed phonetic disintegration (Alajouanine, Ombredane and Durand, 1939), cortical dysarthria (Bay, 1962), apraxic dysarthria (Nathan, 1947), verbal apraxia, apraxia of speech, and dysarthria; however, it is generally agreed that the disorder is one of articulation, secondary to cerebral dysfunction that often co-occurs with aphasia and frequently remains after all, or most, aphasic symptoms are resolved. The present study uses the term "verbal apraxia," defined as impairment of the integrative control of the speech production apparatus due to cortical dysfunction.

---

*Portions of this research were presented at the 1976 Academy of Aphasia Meeting, Miami, Florida.

†Adelphi University, Garden City, N.Y.

††Also Adelphi University, Garden City, N.Y.

†††Also Graduate School, City University of New York.

Acknowledgment: The authors gratefully acknowledge the contributions of their associate Susan Gray-Sweet. We also wish to thank Mr. J.R.P. for his cooperation during many rigorous hours of testing and therapy, and for his faith in our skill as speech pathologists. The research was supported in part by the National Institute of Dental Research DE-01774 to Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

Phoneme productions are characterized by inconsistency, although certain phonemes may never be produced accurately. The oral output is generally slow, labored and dysprosodic. As utilized herein, the term excludes disorders related to bilateral neuromuscular involvement of the speech mechanism.

Research and publication related to this disorder has focused on description and classification, that is, specification of articulatory characteristics of groups of subjects during a single time period (Alajouanine, et al., 1939; Nathan, 1947; Critchley, 1952; Fry, 1959; Bay, 1962; Shankweiler and Harris, 1966; Johns and Darley, 1970; Deal and Darley, 1972; and Martin and Rigrodsky, 1974), and input or sensory deficits (Aten, Johns and Darley, 1971; Rosenbek, Wertz and Darley, 1973). Little is known about the evolution of the disorder and the efficacy of using remedial techniques for its amelioration. It is significant that with the exception of Schuell, Jenkins and Jimenez-Pabon (1964), Rosenbek, Lemme, Ahern, Harris and Wertz (1973); Darley, Aronson and Brown (1975); and Daoul and Bollier (1976), little has been written about specific therapeutic techniques for use with the apractic patient. It has been the empirical observation of aphasiologists that verbal apraxia is a syndrome particularly amenable to treatment, provided that the treatment occurs over a considerable time period--years, rather than months. The present study was undertaken to investigate this observation.

In 1966, Shankweiler and Harris published the results of an experimental procedure that they had devised to analyze and describe disorders of articulation in aphasia. Subjects of their study were five patients who had suffered cerebrovascular accidents. In each case the stroke was followed by right-sided involvement and expressive aphasia with minimal comprehension deficits. Within the early recovery period, most of the aphasic symptoms were resolved, and articulation problems were the major residual. Each patient was given a battery of tests, including a test of speech perception, audiometric evaluation and an articulation assessment.

The articulatory assessment consisted of a test of 200 real word monosyllables. The list contained most singleton consonants, a sample of the most frequently occurring consonant clusters, and those vowels that are not ordinarily characterized by glides in the regional dialect. A major difference between this and all other tests was that each consonant and consonant cluster occurred eight times in each position, while the eight vowels occurred twenty-five times each. In a disorder characterized by inconsistency, this was important in order to draw inferences about the relative difficulty of each phoneme. The patient's task was to repeat each word once. Responses were tape-recorded and later analyzed by a phonetically-trained listener, using broad phonetic transcription. The transcribed utterances were tabulated as confusion matrices in order to illustrate the frequency with which each phoneme and each cluster was correctly produced or replaced by another phoneme or cluster.

Shankweiler and Harris concluded that for four of their five subjects, initial phonemes were more difficult to produce than terminal phonemes. The vocalic portions of the words were produced with greater accuracy than the nonvocalic portions. Four of the five patients demonstrated accurate production of most vowels. Errors on consonants were classified in terms of whether the substituted sound differed from the target phoneme in (1) place of

articulation, (2) manner of articulation, (3) both place and manner, or (4) whether the substituted phonemes were unrelated to the target or omitted completely. The results of the tabulations showed a similar incidence of errors of place and manner, with manner errors slightly predominating. The largest category of misarticulations, accounting for one-third of all errors, was the category of unrelated substitutions and omissions. Particularly common were substitutions of consonant clusters for single consonants, a phenomenon rarely observed in other disorders of speech production.

PROCEDURE AND RESULTS

One of the subjects of the Shankweiler-Harris study (J.P.) continued to seek and receive therapy during the ten-year period since the completion of the study. It therefore seemed worthwhile to replicate the testing procedure of the earlier study and thus derive comparative longitudinal data on this single patient.

J.P. suffered a CVA in 1964. He was a subject for the Shankweiler-Harris study during 1965, approximately six to nine months post-onset. He is now 56 years of age and has been healthy since the initial CVA. He continues to demonstrate right spastic hemiparesis and verbal apraxia as his major residual symptoms. He has received speech therapy, both group and individual, for 2-3 hours per week during the past ten years. Therapy was conducted by a number of clinicians in two rehabilitation settings. In general, the therapeutic techniques followed traditional procedures emphasizing accuracy of articulation in imitation of the clinician. In the fall of 1975, a battery of tests for aphasia indicated only minimal residual deficits in language. Hearing was normal throughout the speech frequencies. On a test of auditory perceptual abilities, the Test of Differentiation of Phonemic Feature Contrasts (Mitchell, 1974), his perception of speech was within the normal range.

The articulatory assessment test devised by Shankweiler and Harris (1966) and previously described, was administered. Figure 1 consists of two sets of bar graphs labeled 1965 and 1975. The 1965 graphs are extrapolated from Shankweiler and Harris (1966), as the results on J.P. The percentages shown are approximations rather than exact percentages, since they were derived from a graph. Vowel production was not a significant problem for J.P. in 1965 nor in 1975, and is therefore omitted from the presentation of results. The differences between 1965 and 1975 are readily apparent. J.P.'s production of initial clusters, initial consonants and terminal consonants improved greatly. Production of terminal clusters was essentially unchanged. Overall, in 1965, 51 initial phonemes, or 25 percent, were correctly produced; while in 1975, 142 initial phonemes, or 71 percent, were correctly produced.

These data are presented in greater detail in the confusion matrices of Figures 2 and 3. In the matrices the target phonemes appear on the vertical axis, and the phonemes produced appear on the horizontal axis. Cells forming the diagonals indicate correct productions.

Some pattern shifts are recognizable on the matrices. For example, in 1965, /d/ was substituted 43 times and /b/ was substituted 30 times; together, /b/ and /d/ substitutions accounted for 49 percent of the 1965 errors. In 1975, /b/ and /d/ were each substituted four times, together constituting only

FIGURE 1

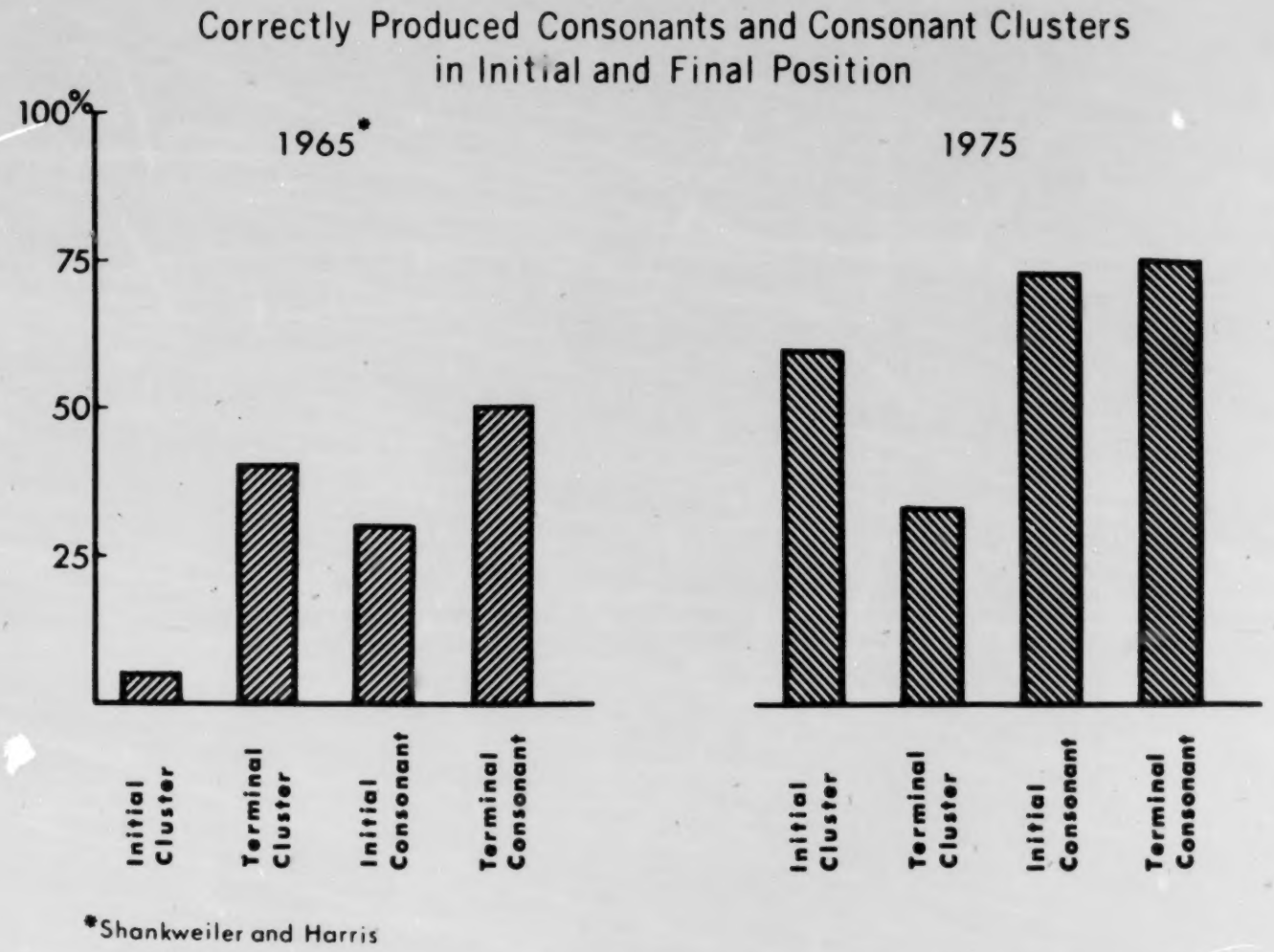

Figure 1: Comparisons (1965 to 1975) of correctly produced consonants and consonant clusters in initial and final position.

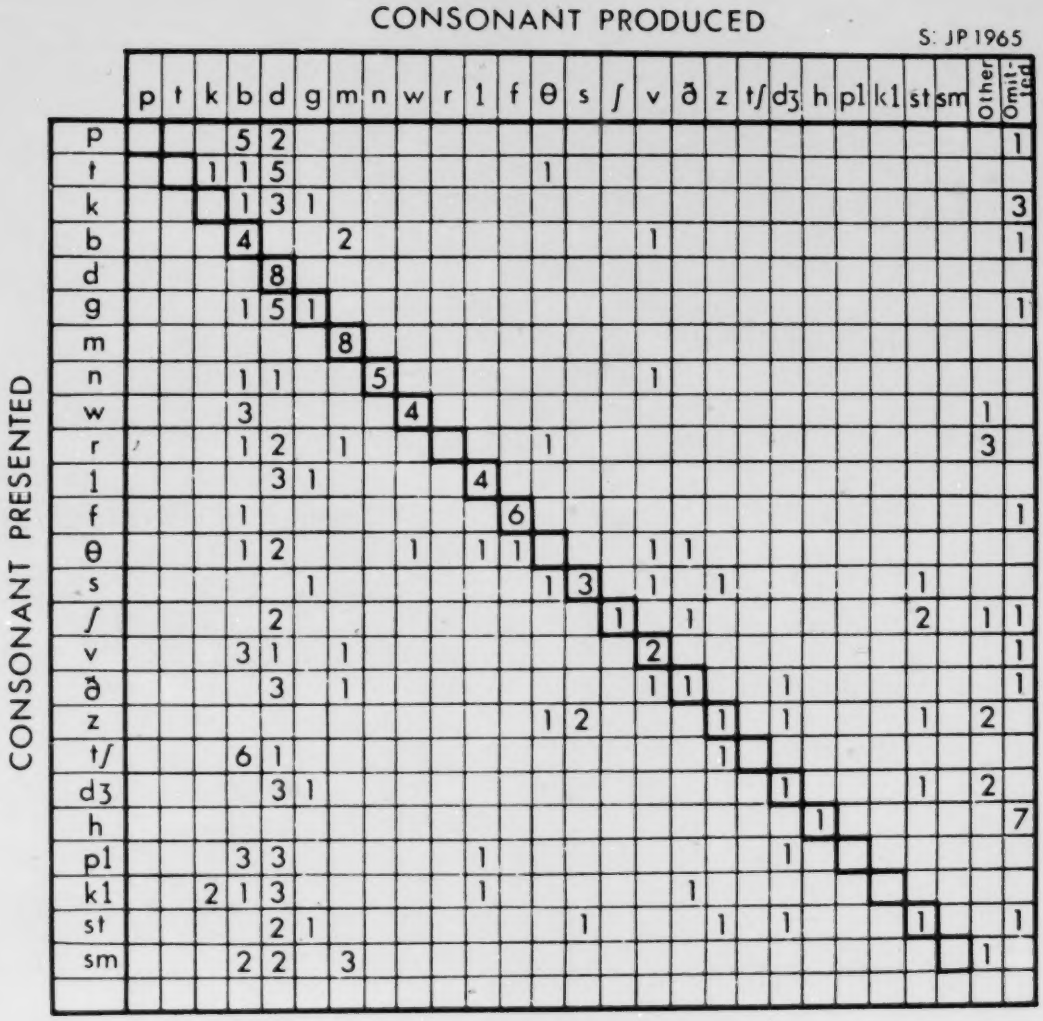
CONSONANT PRODUCED
S: JP 1965
CONSONANT PRESENTED
p t k b d g m n w r l f θ s ʃ v ð z tʃ dʒ h pl kl st sm Other Omitted
p 5 2 1
t 1 1 5 1
k 1 3 1 3
b 4 2 1 1
d 8
g 1 5 1 1
m 8
n 1 1 5 1
w 3 4 1
r 1 2 1 1 3
l 3 1 4
f 1 6 1
θ 1 2 1 1 1 1 1
s 1 1 3 1 1 1 1
ʃ 2 1 1 2 1 1
v 3 1 1 2 1
ð 3 1 1 1 1 1
z 1 2 1 1 1 2
tʃ 6 1 1
dʒ 3 1 1 1 2
h 1 7
pl 3 3 1 1
kl 2 1 3 1 1
st 2 1 1 1 1 1
sm 2 2 3 1

Figure 2: Confusion matrix of 1965 phoneme errors.

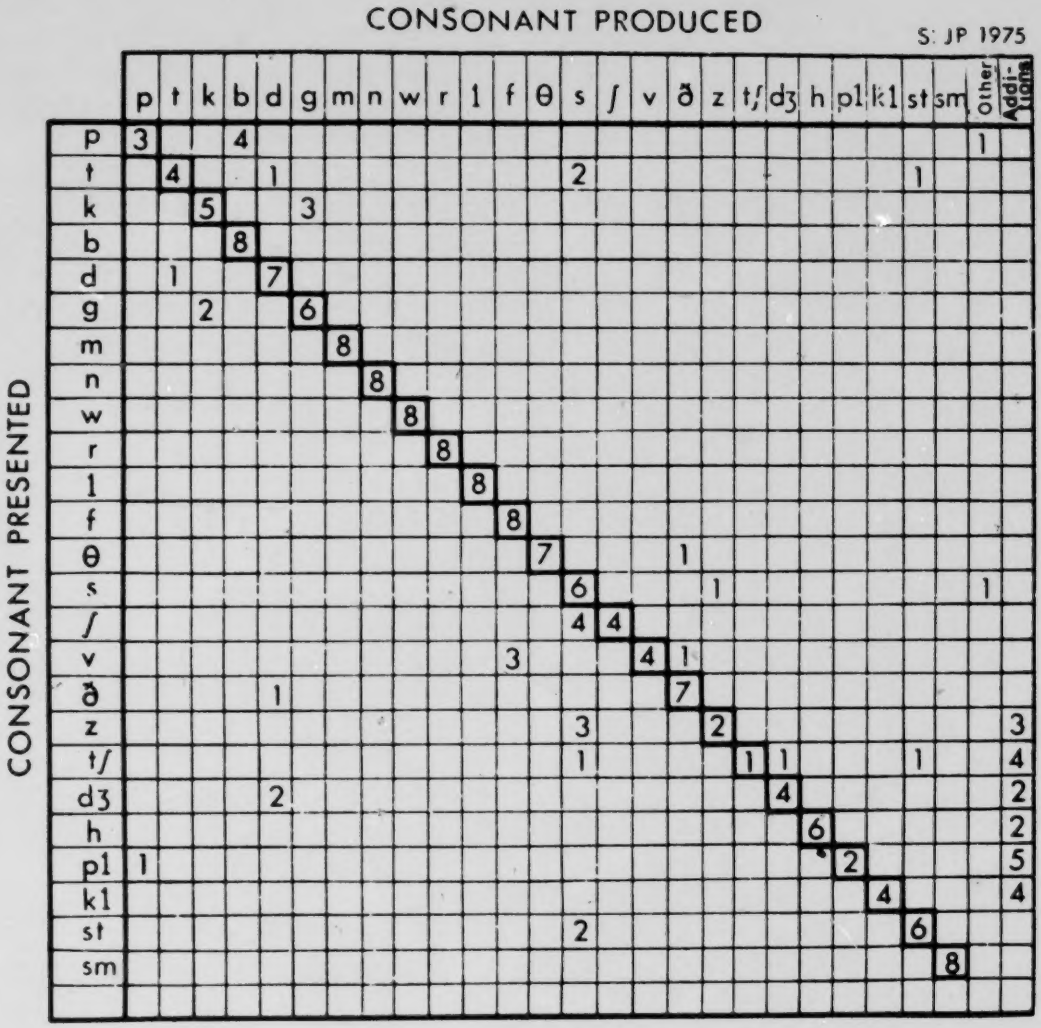

Figure 3: Confusion matrix of 1975 phoneme errors.

14 percent of the errors.

In 1965, 18 omissions were recorded, constituting 18 percent of the total errors. In 1975, no singleton consonants were omitted, indicating a marked pattern change.

In 1965, /s/ was substituted three times, constituting 2 percent of the errors, while /s/ substitution constituted 20 percent of the 1975 errors. In the 1975 matrix, 11 errors involved the addition of /s/ to a consonant or cluster. Errors involving /s/ (substituted or added) constituted 40 percent of the 1975 errors. It is clear that both the number and the nature of the errors changed dramatically during the 10-year period.

Table 1 examines the nature of the errors through a feature analysis.[1] A single misproduced phoneme may differ from the target phoneme by one, two, or more features. Therefore, the error totals (Figures 2 and 3) do not correspond to feature error totals. Shankweiler and Harris (1966) included "voicing" errors as a part of their "manner" category. The present study includes voicing and manner as separate categories. Since 10 of the 1965 errors were classified only as "other," these could not be analyzed by features; therefore, only 139 errors are included in the feature analysis for 1965. The omissions on the 1975 feature analysis are the result of omitting one element in a cluster, and are listed as substitutions on the matrix. For example, /s/ was substituted for /st/ twice. Voicing, place, manner and omission errors were all reduced in total number. The apparent increase in the number of errors of addition may result from our inability to analyze the 10 errors designated as "other" in the 1965 study.

Place and manner errors that were a factor in 31 percent and 37 percent of the 1965 errors were present in only 9 percent of the 1975 errors. At the same time, omission errors decreased from 18 percent to 5 percent. In contrast, voicing errors that were present in 37 percent of the 1965 errors, were present in 34 percent of the 1975 errors. The addition errors that occurred in 34 percent of the 1975 errors were very consistent in nature. These errors all involved the addition of an initial voiceless fricative, usually perceived as /s/ by listeners.

---

[1] A simple, conventional feature analysis was used. Consonants (other than /h/) were considered to be labial, alveolar, palatal or velar in place of production, and stop, fricative, continuant or nasal in manner. Substitutions of /s/ for /θ/ were considered to be errors of place, while /r/-/l/ substitutions were considered as errors of manner only when these particular contrasts were made. Counting of errors of manner, place and voicing for each phoneme in clusters increased the total number of errors in each category, but did not alter the error pattern, and was therefore omitted. With respect to clusters, omissions were scored if one or both members of the cluster were omitted. Errors of voicing, place and manner were scored only for /st/ as noted by dashes in the table. The "additive" category would have been indicated in the 1965 analysis as "other;" phones for which these categories are scored are starred.

TABLE 1: Comparative feature analysis of errors 1965* vs. 1975

	Voicing errors		Place errors		Manner errors		Omissions		Additions	
Phoneme	1965	1975	1965	1975	1965	1975	1965	1975	1965	1975
p	7	4	2	0	0	0	1	0	0	1(s)
t	6	1	2	0	1	2	0	0	0	1(s)
k	5	3	4	0	0	0	3	0	0	0
b	0	0	0	0	3	0	1	0	0	0
d	0	1	0	0	0	0	0	0	0	0
g	0	2	6	0	0	0	1	0	0	0
n	0	0	2	0	3	0	0	0	0	0
w	0	0	0	0	3	0	0	0	*	0
r	1	0	2	0	5	0	0	0	*	0
l	0	0	1	0	4	0	0	0	0	0
f	1	0	0	0	1	0	1	0	0	0
θ	7	1	4	0	5	0	0	0	0	0
s	3	1	3	0	1	0	0	0	1	0
ʃ	3	0	5	4	2	0	1	0	*	0
v	0	3	2	1	5	0	1	0	0	0
ð	0	0	2	0	5	1	0	0	0	0
z	4	3	1	0	1	0	0	0	*	3(s)
tʃ	8	1	6	0	8	0	0	0	0	4(s)
dʃ	1	0	0	0	4	2	0	0	*	2(s)
h	0	0	0	0	0	0	7	0	0	0
pl	-	-	-	-	-	-	1	1	*	5(s)
kl	-	-	-	-	-	-	3	-	*	4(s)
st	5	0	1	0	0	0	2	2	*	0
sm	-	-	-	-	-	-	3	0	*	0
Totals	51	20	43	5	57	5	25	3	1	20
	37%	34%	31%	9%	37%	9%	18%	5%	.5%	34%

*Ten of the 1965 errors were recorded as "other" and could not be analyzed by features. Therefore all percentages for 1965 are based on the 139 errors that could be analyzed.

In summary, between 1965 and 1975, errors of place, manner and omission were markedly reduced, and voicing errors, while reduced in total number, still constituted a substantial percentage of J.P.'s residual errors.

DISCUSSION

On the basis of the analyses performed, it appears that therapy with J.P. had been effective in improving place and manner productions, and in virtually eliminating omission errors. Therapy had apparently been less efficient in dealing with voicing and addition errors. In an attempt to understand the changes that had occurred, the results of the feature analyses were considered in terms of the dynamics of production.

As discussed by Lisker and Abramson (1964, 1967), the voicing contrast in stops is directly linked to the acoustic consequences of the time between the release of the stop occlusion and the initiation of vocal fold vibration. This temporal relationship is linked to the coordination of abductory and adductory forces at the larynx with upper articulator events (Hirose and Gay, 1972). From the feature analysis of J.P.'s errors, it was hypothesized that he was unable to control the initiation of voicing in relation to his stop release.

The addition of initial voiceless friction to a consonant or cluster occurs if airflow commences before the appropriate articulatory or phonatory constriction is achieved. In this paradigm, J.P.'s addition errors may be considered to result from poor temporal control of airflow, vocal fold adduction and/or articulatory gesture.

The results of this analysis suggested that the voicing and addition errors arise from a common defect, that is, poor temporal coordination of airflow, phonation and articulation. Most of J.P.'s residual errors would thus be considered errors of timing.

Three explanations for our subject's pattern of improvement appear possible. First, coordination of airflow, phonation and articulation may be the most difficult task for the apraxic, or for J.P., and therefore this problem remains after the others have been resolved. Second, it is possible that therapeutic procedures that effectively stress and teach place and manner, resulted in correction of these features, whereas therapy was less effective in attacking and correcting (or even recognizing) the effects of incoordination of airflow, phonation and articulation. Finally, it appears possible that a combination of degree of difficulty and therapeutic emphasis resulted in this pattern of performance.

The results indicate that longitudinal studies of apraxtic patients, using systems that allow comparison of consistency of articulation and permit an analysis of the error features, provide critical information about the nature of articulatory dysfunction secondary to cerebral vascular insults. Future studies should test the hypothesis that deficits in coordination underlie a significant portion of the phonemic errors that constitute verbal apraxia.

REFERENCES

Alajouanine, T., A. Ombredane and M. Durand. (1939) Le Syndrome de Disintegration Phonetique dans l'Aphasie. (Paris: Masson).

Aten, J. L., F. L. Darley, J. L. Deal and D. Johns. (1975) Letter to the editor. J. Sp. Hear. Dis. 40, 416-420.

Aten, J. L., D. F. Johns and F. L. Darley. (1971) Auditory perception of sequenced words in apraxia of speech. J. Sp. Hear. Res. 14, 131-143.

Bay, E. (1962) Aphasia and non-verbal disorders of language. Brain 84, 412-426.

Critchley, M. (1952) Articulatory defects in aphasia. J. Laryng. Otol. 66, 1-17.

Dabul, B. and B. Bollier. (1976) Therapeutic approaches to apraxia. J. Sp. Hear. Dis. 41, 268-276.

Darley, F. L., A. E. Aronson and J. R. Brown. (1975) Motor Speech Disorders. (Philadelphia: Saunders).

Deal, J. L. and F. L. Darley. (1972) The influence of linguistic and situational variables on phonemic accuracy in apraxia of speech. J. Sp. Hear. Res. 15, 639-653.

Fry, D. B. (1959) Phonemic substitutions in an aphasic patient. Lang. Sp. 2, 52-61.

Hirose, H. and T. Gay. (1972) The activity of the intrinsic laryngeal muscles in voicing control. Phonetica 25, 140-164.

Johns, D. F. and F. L. Darley. (1970) Phonemic variability in apraxia of speech. J. Sp. Hear. Res. 13, 556-583.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Sp. 10, 1-28.

Martin, A. D. (1974) Some objections to the term "apraxia of speech." J. Sp. Hear. Dis. 39, 53-64.

Martin, A. D. and S. Rigrodsky. (1974) An investigation of phonological impairment in aphasia, part 2: Distinctive feature analysis of phonemic commutation errors in aphasia. Cortex 10, 329-346.

Mitchell, P. D. (1974) Test of differentiation of phonemic feature contrasts. Unpublished Ph.D. dissertation, City University of New York.

Nathan, P. W. (1947) Facial apraxia and apraxic dysarthria. Brain 70, 449-478.

Rosenbek, J. C., R. T. Wertz and F. L. Darley. (1973) Oral sensation and perception in apraxia of speech and aphasia. J. Sp. Hear. Dis. 16, 22-36.

Rosenbek, J. C., M. L. Lemme, M. B. Ahern, E. Harris and R. T. Wertz. (1973) A treatment for apraxia of speech in adults. J. Sp. Hear. Dis. 38, 462-472.

Schuell, H., J. J. Jenkins and E. Jimenez-Pabon. (1964) Aphasia in Adults. (New York: Harper & Row).

Shankweiler, D. and K. S. Harris. (1966) An experimental approach to the problems of articulation in aphasia. Cortex 2, 277-292.

Temporal Coordination of Phonation and Articulation in a Case of Verbal Apraxia: A Voice Onset Time Study*

Frances J. Freeman,† Elaine S. Sands†† and Katherine S. Harris†††

ABSTRACT

A study of voice onset time (VOT) in stop production was undertaken in order to investigate the hypothesis that the voicing feature errors in the speech of an apraxic patient (Sands, Freeman and Harris, 1977) were related to deficits in temporal coordination of phonation and articulation. Results demonstrated that the VOTs of the apraxic subject differed markedly from those of normal subjects. The apraxic productions did not include voicing lead for voiced stops. Lag times for voiced stops were longer than normal, while those for voiceless stops were shorter than normal, yielding a compression of the two categories and a marked overlap.

INTRODUCTION

In our discussion of the results of an analysis of progressive changes in articulatory patterns in verbal apraxia, Sands, Freeman and Harris (1977) hypothesized that the voicing errors that occurred in 34 percent of our subject's residual errors resulted from defective temporal coordination of phonation and articulation. A large portion of the subject's voicing errors occurred on stops. Thirty-three percent of his initial voiceless stop productions were perceived by listeners as being voiced.

The present study proposed to test the temporal coordination hypothesis in relation to stop production.

---

*Portions of this research were presented at the 1976 Academy of Aphasia, Miami, Florida.

†Also Adelphi University, Garden City, N.Y.

††Adelphi University, Garden City, N.Y.

†††Also Graduate School, City University of New York.

Acknowledgment: The authors gratefully acknowledge the contributions of their associate Susan Gray-Sweet. We also wish to thank Mr. J.R.P. for his cooperation during many rigorous hours of testing and therapy, and for his faith in our skill as speech pathologists. The research was supported in part by the National Institute of Dental Research Grant DE-01774 to Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

PROCEDURE AND RESULTS

In order to test the hypothesis that J.P.'s residual problem was one of coordination of upper articulatory and laryngeal events, voice onset time (VOT) was measured for initial stops. The procedures essentially followed those used in a classic study of voicing in stops by Lisker and Abramson (1964; 1967). The subjects read a randomized list of words beginning with the six English stops (p/b, t/d, k/g). Wide-band spectrograms were made from the recordings. On spectrograms, the oral release of stop occlusion is marked by an abrupt change of acoustic energy, while glottal pulsing is indicated by regularly spaced vertical striations. The time between the release of occlusion and onset of voicing can thus be measured. By convention, the burst is considered as zero time. Glottal pulsing preceding the burst (voicing lead) is measured and expressed as a negative number, while glottal pulsing following the burst (voicing lag) is measured and expressed as a positive number. In the apraxic production, there was some instability in voicing, not often found in normal productions. Occasionally, a few pulses of voicing would occur, and then voicing would cease, to begin again later in the production. For this reason the first pulse of "continuous" voicing was measured.

The results are presented in Figures 1, 2 and 3. In each of these figures the top graph presents data on J.P., while the middle and lower graphs present the normative VOT data of Lisker and Abramson (1967).

In their study, Lisker and Abramson (1967) examined VOT for syllable initial stops in both isolated words and words in sentence context. The middle and lower graphs of Figure 1 illustrate their findings for the bilabials /b/ and /p/. The temporal categories for /p/ and /b/ are relatively discrete with an overlap in the sentence condition only. The /b/ in isolated words ranges from a long lead of -130 msec to a brief lag. In sentences the normals shortened their lead time and occasionally showed a brief voicing lag. Means were calculated separately for voicing lead and voicing lag. (Since J.P. always showed voicing lag, the mean lag times for normals are marked on the figures with arrows.) The /p/ in isolated words ranges from +20 msec to +120 msec, with a mean of +59 msec (indicated by an arrow in the middle graph). Data for /p/ and /b/ in sentence context are presented in the same format in the bottom graph. The mean lead times for /b/ and /p/ differ by 57 msec for the isolated words and by 28 msec for the words in sentences.

In contrast, J.P. never used a voicing lead. His ranges for both /p/ and /b/ are markedly compressed and show significant overlap. His means for /p/ and /b/ (indicated by arrows in the top graph) differ by only 6 msec. When a line is drawn at the intersection of the VOT distributions for J.P., 70 percent of his productions fall to the left of this line and only 30 percent fall to the right.

Figure 2 presents data for the alveolar stops in parallel format. Normals again show relatively discrete temporal categories for production of these phonemes. For normals the mean lag times differ by over 60 msec for isolated words, and by over 30 msec for words in sentences. For /t/ and /d/, J.P.'s categories were again compressed with a difference between means of only 14.5 msec.

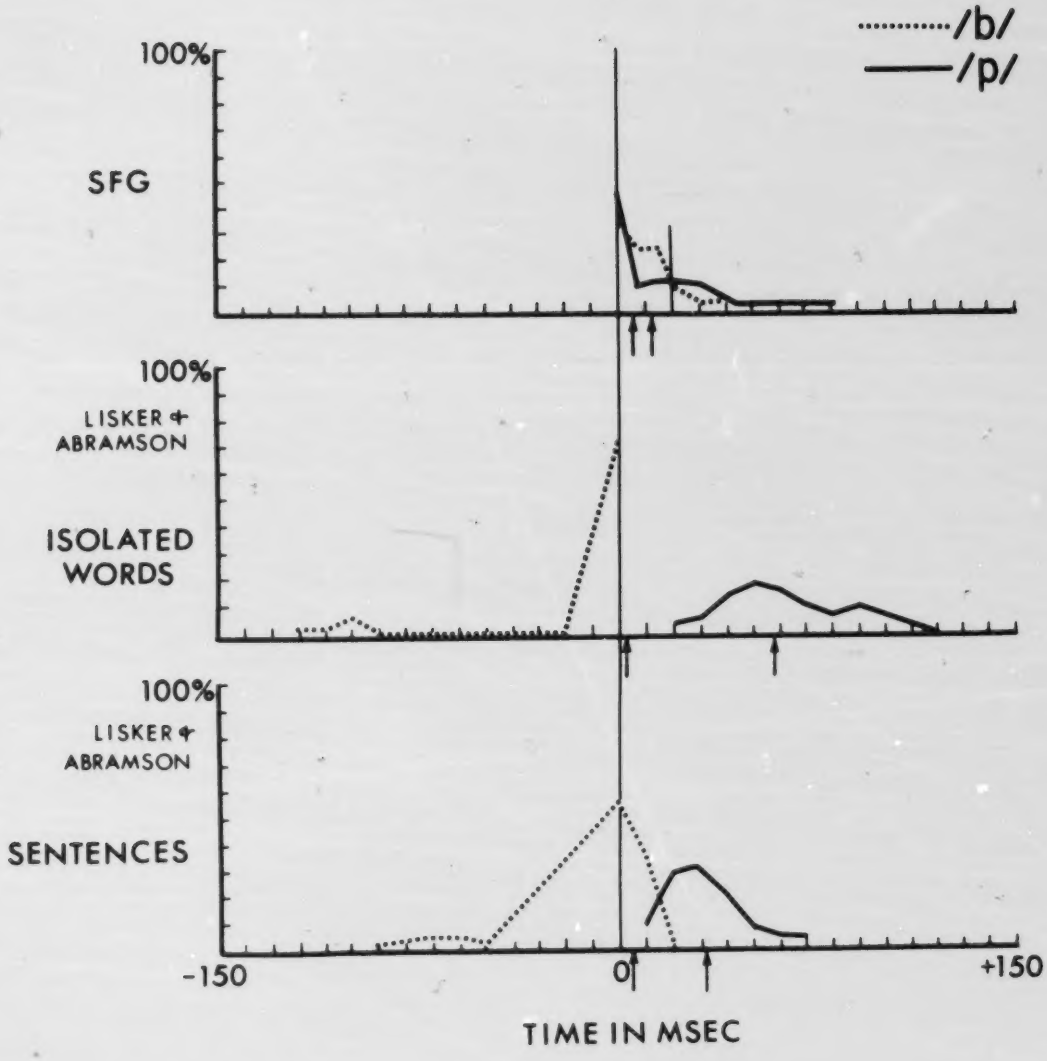

Figure 1: Comparison of apraxic and normal VOT for bilabial stops.

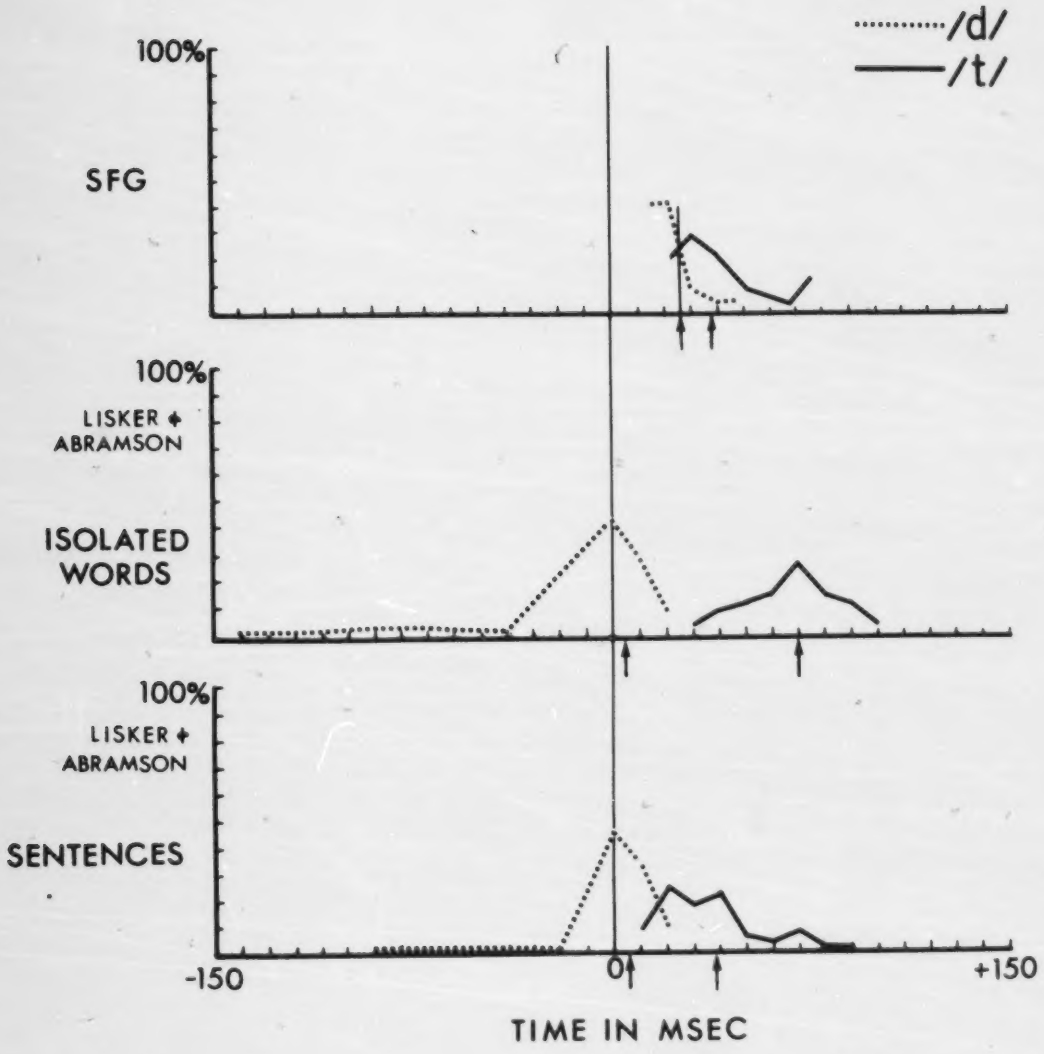

Figure 2: Comparison of apraxic and normal VOT for alveolar stops.

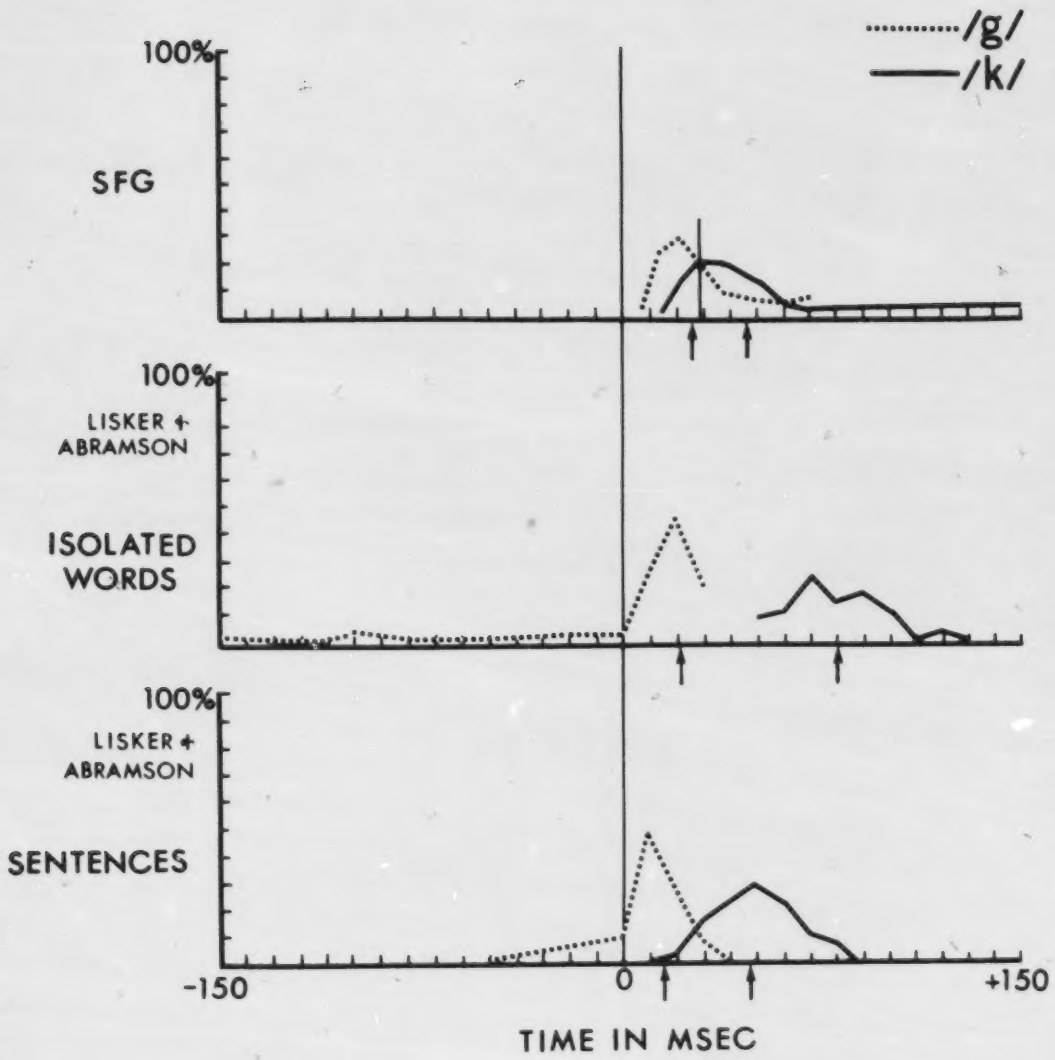

Figure 3: Comparison of apraxic and normal VOT for velar stops.

The pattern for velar stops /k/ and /g/ (Figure 3) is very similar. J.P. shows compression of the categories, with the mean for /g/ and the mean for /k/ differing by only 21 msec. Again there is a marked overlap of the two categories.

DISCUSSION

After analysis of the stop productions was completed, these results were compared with the listener perceptions of J.P.'s stops on the articulatory assessment administered in the earlier study (Sands, Freeman and Harris, 1977).

In the VOT study it was found that 30 percent of J.P.'s /p/ productions fell to the right of the intersect line, while in the articulation study, 37.5 percent of his /p/ productions were perceived as being /p/. Seventy percent of his /p/ productions had shorter lag times and fell to the left of the intersect line, and 50 percent of his /p/ productions were perceived as /b/. Despite the use of different sets of utterances, there appears to be a relationship between the subject's VOT and listener perceptions.

The category overlap for his /k/ and /g/ productions also appears to relate directly to listener perceptions. Listeners in the articulation study perceived 75 percent of J.P.'s /g/ productions correctly, while perceiving 25 percent of his /g/ productions as /k/. On the VOT study, 78 percent of J.P.'s /g/ productions fall to the left of the intersect line, while 22 percent are to the right. A similar relationship between production and perception exists for /k/. Listeners in the articulation study perceived 37.5 percent of J.P.'s /k/ productions as being voiced and 62.5 percent as being voiceless. Since 36 percent of J.P.'s /k/ productions fall to the left of the intersect line, while 64 percent fall to the right, the intersect line is a highly accurate division of what listeners perceive as /k/ and /g/. The alveolar stops could not be compared in this manner because errors in the articulatory study included manner and place as well as voicing.

There is a well-known hypothesis, expressed by Jakobson (1968), among others, that the defect found in apractic articulation can be considered a regression to a more primitive form of articulation--that of young children. Fortunately, VOTs have been extensively studied in young children, in their earliest word productions, and in babbled utterances containing stop-like articulations (Kewley-Port and Preston, 1974). They show that the earliest examples of apical stops, produced around 6 months of age, have uniform distributions along the VOT continuum. Later, the distribution of stops collapses to an interval corresponding to that of American English voiced stops. When recognizable words are first produced, the range of "d" words is quite like short-lag /d/ production in adults. However, the earliest /t/ words are produced with a voicing lag which would be ambiguously categorized as /t/ or /d/. The VOT distributions produced by J.P. are quite similar to children in this developmental stage--that is, productions have a short lag.

Voice onset time studies of other apraxic speakers are necessary to determine whether this problem is idiosyncratic or a common feature of the disorder. Studies of other temporal parameters, such as fricative and vowel duration, will be necessary to the further understanding of temporal coordina-

tion deficits in apraxia. Research in these areas has been undertaken and preliminary results reported (Jensen, MacDonald and McGurk, 1975; Freeman, Gray and Sands, 1976). Finally, the authors, as clinicians, were interested in the applicability of these findings to the development of therapeutic strategies. Results of a therapeutic program for teaching voicing lead and lag to J.P. (Leavitt, Sands and Freeman, 1977) are encouraging.

REFERENCES

Freeman, F. J., S. M. Gray and E. S. Sands. (1976) Disruption of temporal integration in apractic speech production. Paper presented at the annual convention of the American Speech and Hearing Association, Houston, Texas.

Jakobson, R. (1968) Child Language, Aphasia, and Phonological Universals. Translated by A. R. Keller. (The Hague: Mouton).

Jensen, T. W., P. L. MacDonald and M. P. McGurk. (1975) Peak intraoral air pressure, duration of intraoral air pressure, and voice onset time in aphasia of speech. Paper presented at the annual convention of the American Speech and Hearing Association, Washington, D.C.

Kewley-Port, D. and M. S. Preston. (1974) Early apical stop production: A voice onset time analysis. J. Phonetics 2, 195-210.

Leavitt, N., E. S. Sands and F. J. Freeman. (1977) Use of voice onset time data in the design of therapy for apraxia of speech. Paper presented at the New York State Speech and Hearing Association Meeting, Rochester, New York.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Sp. 10, 1-28.

Sands, E. S., F. J. Freeman and K. S. Harris. (1977) Progressive changes in articulatory patterns in verbal apraxia: A longitudinal case study. Haskins Laboratories Status Report on Speech Research SR-51/52.

Factors in the Maintenance and Cessation of Voicing*

Leigh Lisker†

ABSTRACT

There is wide agreement as to the articulatory and acoustic facts of voicing: approximation of the vocal folds and establishment of a difference in the air pressures above and below the glottis. When these conditions prevail and airflow through the mouth and nose is blocked, there is some uncertainty as to what other factors determine whether vocal fold vibration is maintained or extinguished. One view is that special maneuvers are required to ensure maintenance of vibration--readjustment of the larynx so as to reduce glottal resistance to airflow and enlargement of the supraglottal cavity. Another view, not so often expressed now, is that voicelessness requires tensing of the supraglottal musculature so as to prevent expansion of that cavity as supraglottal air pressure builds up behind the articulatory closure. The first view implies that stop consonants are "naturally" voiceless; the second would seem to imply the opposite, unless one assumes tenseness to be the normal state of the supraglottal tract, in that it says the surfaces enclosing the supraglottal cavity will move possibly in response to the buildup of airpressure so as to permit air to continue flowing through the glottis during occlusion. Thus the explanatory literature on stop voicing is a rich one; in fact it is too rich, for it permits us to suppose that the contrast may be independent of any differential laryngeal adjustment at all, when in fact the evidence continues to mount that the larynx is, as we have known all along, actively implicated.

In preparing a comment for this meeting I found myself wondering just why the larynx seems to attract so much more attention than any of the other parts of the speech-producing apparatus. Not that we have no questions about the functioning of the velum, tongue, jaw, lips and respiratory musculature; but the larynx is especially provocative of question and debate. Of course we are all in agreement about some things: the larynx is the source of the quasi-periodic signal which characterizes most speech that we call "voiced," and a glottal airflow is a necessary though not sufficient condition for the

---

*This short discussion was prepared for a seminar on The Larynx and Language held at the Eighth International Congress of Phonetic Sciences, Leeds, England, 17-23 August 1975. A version of this paper will appear in Phonetica.

†Also University of Pennsylvania.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

oscillatory movement of the vocal folds that provides the modulation of this airflow. Moreover, if we limit our attention to voicing as a distinctive feature of the languages best known to Europeans, then we can believe, erroneously of course, that the larynx is a two-state device so far as its linguistic function is concerned: the folds either vibrate at an audible frequency or they do not. In this view the different kinds of vibratory patterns that the folds can execute are equally [+voice]; indeed, one of them may be called the "normal" mode of vibration, all the others being relegated either to paralinguistic function or to certain exotic languages we are inclined to ignore, except as we are taught better by colleagues like Eugénie Henderson.[1] Can we similarly presume to suppose that there is a state of the larynx that is "normal" for speech characterized by the absence of vocal-fold vibration? The answer seems to be "no": the absence of modulation of a glottal airflow does not suffice as unambiguous evidence for the state of laryngeal adjustment. The folds may be approximated or tightly closed, or they may be widely separated. If we are certain that there is a glottal airflow and that it is not modulated by vocal-fold oscillation, then we are somewhat surer that the folds are well separated. Just how much they are separated is perhaps related, as Kenneth Stevens[1] suggests, to other factors that may be operating to prevent the passing air from setting up appreciable oscillation of the folds. But it must be noted that if the voicelessness of particular consonants is said to involve, necessarily or even optionally, some action to stiffen the folds, we are so far without observational data to support this view.

The straightforward picture of the larynx as an on-off generator gives way to a more complex situation when we consider a fact of language--namely, that voicing as a distinctive property of the speech signal most often occurs in conjunction with a severe constriction of the oral cavity. For the condition in which air flows through the glottis, there may not be a unique state of the larynx for "oscillator-on" and another for "oscillator-off" operation, but we may be certain that in switching from one condition to the other, the larynx itself undergoes some adjustment.

However, if there is blockage of the airflow somewhere above the larynx, an interruption of fold vibration need not, on the face of it, be ascribed to any particular laryngeal adjustment; it might be simply a consequence of oral occlusion maintained long enough to halt the glottal airflow needed to support oscillation. If voiceless stops tend to have closures of longer duration than voiced, this would be consistent with a view that articulatory closure is at the same time a devoicing gesture. Calculations reported by Morris Halle and Kenneth Stevens (1967) and by Martin Rothenberg (1968) suggest, however, that the durations of even the voiced stops are greater than would seem needed to halt glottal airflow. This raises a question to which we are still without a sure answer: How is this airflow maintained during the oral closures of the order of 100 msec that are not unusual in speech? Observational data are reported that show enlargement of the supraglottal cavity during voiced stop production: velar elevation, tongue advancement, larynx lowering, and each of these maneuvers should reduce the pressure-equalizing effect of the oral

---

[1]Colleagues at the Congress.

closure. But whether cavity-enlarging maneuvers and passive response of the cavity walls to airpressure change, assuming the cavity is tightly sealed, are sufficient to account for the durations of observed voiced-closure intervals, is still not entirely clear.

It is, in fact, not all that certain that the various cavity-enlarging maneuvers available to the speech mechanism are regularly performed during voiced stop production; thus, for example, Perkell's (1969) well-known X-ray analysis reports that "there is little observable effect of the different consonants on the behavior of vertical movement of the hyoid bone and larynx." In any case, however, the fact that voicing may persist unbroken through an interval of oral closure has elicited an explanatory literature. On the other hand, it has not been generally agreed that voiceless closure intervals require no devoicing maneuver other than the articulatory closure itself. Thus the voiceless stops (and fricatives also) have been on occasion described as involving a feature of general tenseness, a tensing of the supraglottal cavity walls, and/or a more specifically laryngeal tensing.

Given the extensive literature on stop voicing, one is entitled to ask whether it is the voicing or the devoicing of a closure interval that forces us to invoke a maneuver or maneuvers over and above the aerodynamic effect of the unaided closure. Of course we might also suppose that the question is poorly posed in "either-or" terms: to ensure voicing could require cavity enlargement or oral leakage, while reliable devoicing might necessarily involve positive prevention of any such enlargement, perhaps even a contraction of the supraglottal volume. If this were true in fact, then we should be greatly tempted to believe seriously that the stop voicing distinction need involve no specifically laryngeal adjustment. Whether or not such an adjustment is strictly necessary, however, it is an incontestable fact that in the production of voiceless stops there is clear EMG evidence of contraction of the posterior cricoarytenoid muscles. This action, if it is redundant, is evidence that the notion of economy of articulatory effort can be taken too seriously as an explanation of speech phenomena. But perhaps it is risky to write off the activity of the posterior cricoarytenoids as an instance of the "unmotivated" expenditure of articulatory effort. What is most certain in all this is that stop voicing will continue to provide problems to exercise us, assuredly until the next international congress.

REFERENCES

Halle, M. and K. N. Stevens. (1967) On the mechanism of glottal vibration for vowels and consonants. Quarterly Progress Report 85. (Research Laboratory of Electronics, MIT), 267-271.

Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. (Cambridge: MIT Press).

Rothenberg, M. (1968) The breath-stream dynamics of simple-released-plosive production. Bibliotheca Phonet. no.6. (Basel: Karger).

The Influence of Tempo on Stop Closure Duration as a Cue for Voicing and Place

Robert F. Port[†]

ABSTRACT

The closure interval of poststress medial stops varies in duration due to phonological voicing and place of articulation, as well as speaking tempo. In this experiment the closure duration of the medial stop in rabid was varied over a range of 10 to 120 msec with no glottal pulsing during the closure. Embedded after fast and slow carrier sentences, the variants were identified as ratted, rabid, or rapid. The locus of the perceptual boundary between /b/ and flap at about 30 msec was not significantly influenced by the tempo of the carrier, while the locus of the /b/-/p/ boundary at 75 msec moved nearly 10 msec toward shorter values in the fast carrier sentence. These results, interpreted in terms of constraints on speech production, were taken to imply that closure duration as a voicing cue must be defined relative to speaking tempo. The flap effect suggests that closure duration can sometimes be a decisive cue to place of articulation capable of overriding even naturally produced spectral information.

INTRODUCTION

Although many factors are known to influence jointly the timing of speech events in production (Klatt, 1976), relatively little is known about the interaction of these timing factors in speech perception. Production measurements have shown, for example, that the duration of articulatory closure corresponding to medial poststress English stops varies systematically with speaking tempo (Gaitenby, 1965; Port, 1976), stop voicing (Lisker, 1957; Port, 1976) and place of articulation (Sharf, 1962; Port, 1976). Perceptual studies have demonstrated that closure duration can be an effective cue for stop

---

[†]Indiana University, Bloomington.

Acknowledgment: The main experiment described in this paper is drawn from my 1976 University of Connecticut doctoral dissertation and was presented at the 91st Meeting of the Acoustical Society of America, held in Washington, D.C., April, 1976. I am grateful to my doctoral committee, Arthur S. Abramson, Ignatius G. Mattingly and Alvin M. Liberman, for contributions to the earlier version of this study. David B. Pisoni, Diane Kewley-Port and Arthur S. Abramson made helpful comments on a draft of this paper. I also thank the staff of the Speech Department Listening Room at Brooklyn College, C.U.N.Y. for help in running subjects, and Rosemarie Rotunno for assistance in collecting data for the dababba experiment.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

voicing (Lisker, 1957) and have suggested its sensitivity to the speaking tempo of the context (Pickett and Decker, 1960). One purpose of the present experiment was to investigate the phonetic cue value of stop closure duration under different conditions of speaking tempo.

In spectrographic measurements Lisker (1957) found that the duration of articulatory closure of intervocalic labial stops in English was shorter for the voiced stop, as in rabid, than in the voiceless stop, as in rapid (in addition to differences in the duration of the preceding vowel, formant slopes and the proportion of closure that showed glottal pulsing). In a perceptual experiment employing tape splicing, Lisker found that after all glottal pulsing was removed from the closure, subjects' perception of a test word containing medial /b/ or /p/ could be controlled by manipulating the duration of the closure interval. Thus, judgments of phonological [+ voice] (or for some phonologists, [- tense]) were not entirely predictable from the presence or absence of phonetic voicing (glottal pulsing) during the closure interval. Moreover, if the closure interval was silent, the duration of the closure was a very powerful cue to the voicing feature of the stop.

It has also been found that certain phonologically irrelevant factors such as speaking tempo will influence the durations of intervals in production (Gaitenby, 1965; Kozhevnikov and Chistovich, 1965; Peterson and Lehiste, 1965). A previous study (Port, 1976) found that while there were consistent durational differences attributable to voicing and place between /b,g/ and /p,k/ at a given speaking tempo, changes in speaking tempo also produced large effects on the duration of stop closure. It was also observed that the durations of intervals corresponding to portions of the sentence-stressed word were affected by smaller proportions under tempo change than were unstressed portions of the carrier sentence (Peterson and Lehiste, 1965). Thus, although an increase in speaking tempo will ordinarily cause a decrease in the durations of acoustic intervals, the amount of effect on a given interval may often result from an interaction between tempo and other suprasegmental variables (Gay, Ushijima, Hirose and Cooper, 1974; Klatt, 1976).

On the basis of these production results, one would predict that the perceptual boundary along a continuum of silent closure intervals would not be absolute, but would change depending on the speaking tempo of a carrier sentence. Although we should be able to predict the direction of change (toward shorter values in a faster tempo carrier), we cannot predict the amount of change. In the present experiment, an attempt was made to replicate the portion of Lisker's results that showed that a medial stop can be perceived as voiced with no glottal pulsing during the closure, and that lengthening of the closure interval will produce the perception of voicelessness. In addition, the sensitivity of this perceptual boundary to the speaking tempo of a preceding carrier sentence was also examined.

With respect to a tempo effect on durational phonetic cues, only two groups of studies have been reported in the literature. First, Pickett and Decker (1960) investigated the duration of stop closure as a cue for two linguistic features quite different from those we investigated here. They began with a natural speech production of the sentence "He was the topic of the year" with stress, aspiration and intonation which were sufficiently neutral so that, when the /p/ closure was lengthened in small steps, the

sentence could be heard eventually as "He was the top pick of the year." The perceptual boundary between a word-medial stop and word-final stop + word boundary + word-initial stop was examined at a range of speaking tempos for the carrier sentence surrounding the test word topic. Pickett and Decker found that this boundary shifted systematically with tempo.

Second, Summerfield and Haggard (1972) and Summerfield (1974a, 1974b, 1975) have found a significant effect of the syllabic rate of a preceding carrier sentence on the voice-onset-time (VOT) boundary between initial voiced and voiceless stops at all three places of articulation. The perceptual shift obtained in their studies was generally in the direction predicted by production measurements; at faster tempos the boundary shifted toward shorter values of VOT. In the following experiment we extended the tempo results found for VOT to the closure duration cue for the voicing feature of medial stops.

Another purpose of this investigation was to study the possible masking of stop formant transitions by following formant transitions. A number of investigators (Abbs, 1971; Fujimura, 1975; Dorman, Raphael, Liberman and Repp, 1975) have found that if a synthetic VC syllable is sufficiently close to a following CV syllable, such as [ɛb] and [dɛ], subjects report hearing only the second stop. That is, [ɛbdɛ] is perceived as [ɛdɛ] when the interval between the vocalic transitions namely, the closure interval, is shorter than 50-80 msec. Similarly, [ɛgbɛ] becomes [ɛbɛ]. Massaro (1972, 1974) has interpreted this effect as evidence of backward masking of the formant transitions on the first imploding stop by the exploding transitions on the second stop. The second set of transitions is assumed to interrupt the processing of the first stop before a perceptual decision has been achieved. Thus the subject is able to perceive only the place of articulation of the second unmasked stop due to constraints in the auditory system. Dorman et al. (1975) have argued, however, that this result is primarily a phonetic effect. They found that the isolated F_2 transitions from imploding [ɛg] and [ɛb], which are not heard as speech but as nonspeech "chirps," were not masked by a following [dɛ], even though they could be assumed to require the same transition processor to be discriminated. On the other hand, if the (synthetic) voice speaking the masker [dɛ] was modified so that subjects heard it as that of a different speaker (for example, as female rather than male), there was again no masking of the imploding stops. Thus they conclude that the effect is not auditory, but rather results from listeners attending to the stimuli as speech sounds from the vocal tract of some human being. In this view, the failure to identify one of the stops reflects constraints on phonetic timing due ultimately to production limitations, since a sequence of two stops produced by the same talker would require a much longer closure interval than a single stop. When the closure interval is too short for a sequence of two stops, subjects are faced with contradictory information about the place of a single consonant and generally perceive the second stop because, presumably, the place cues are more distinctive on explosion than on implosion (Sharf and Hemeyer, 1972; Repp, 1976). In short, the duration of stop closure is, like formant transitions and bursts, simply information about the kind of articulatory gesture the speaker produced. In the words of Dorman et al. (1975), phonetic perception is constrained "not by what the auditory system can do but by what vocal tracts can do."

If closure duration is a major cue to the number of stops between vowels, it might also serve as a distinctive cue to a particular place of articulation. The apical flap (or tap) is, in fact, far shorter than stops at other places of articulation--about 10-30 msec compared to 40-80 msec for intervocalic /b/ and /g/ (Sharf, 1962; Port, 1976). If this durational difference could be used to specify apical place of articulation despite spectral information for some other place on both sides of the stop closure, it would provide support for the Dorman et al. (1975) conclusion that the closure interval specifies phonetic information rather than simply providing an interstimulus interval sufficient to evade backward masking.

In the following experiment, a continuum of silent closure durations was constructed for the medial stop in a naturally produced rabid from the range of closures characteristic of apical flaps through those characteristic of /b/, out to those of /p/. These stimuli were appended to two carrier sentences, one produced at a fast speaking tempo and one at a slow rate, and then presented to subjects for identification as ratted, rabid or rapid.

METHOD

The sentence "I'm trying to say rabid" was recorded by the experimenter at a carefully enunciated slow tempo and at a very fast rate. The signals were digitized (and low-pass filtered at 4 kHz on the Haskins Laboratories pulse-code-modulation (PCM) system for computer-controlled cutting and splicing (Cooper and Mattingly, 1969). The word rabid was removed from both utterances of the sentence at a zero crossing in the waveform immediately after the maximum depression of F_3 for the [r]. The remaining utterances became the fast and slow carrier sentences. The slow version of the word rabid was then modified in the following ways. Since the initial [r] sounded abnormally long when rabid was combined with the fast carrier sentence, the first 100 msec were removed. Then the [ræb] syllable was isolated by cutting immediately after the [b] closure and [bəd] was separated by cutting just before the [b] release. Heard separately these utterances sounded like [ræb], with an unreleased final stop, and [bəd]. These two syllables, originally separated by a 70 msec voiced closure period, were then recombined, separated by varying intervals of silence from 10 msec to 120 msec, to create twelve different versions of the test word. These twelve test words were then appended to copies of both carrier sentences. With test words attached, the fast sentences averaged about 70 percent of the duration of the slow sentences. Separate fast and slow stimulus tapes were prepared for the listening tests. Each contained ten copies of each of the twelve test sentences in randomized order. Trials were separated by three second pauses. Ten practice trials were prefixed to the beginning of each tape.

Sixteen students, enrolled in an introductory phonetics course at Brooklyn College, listened to the tapes binaurally over earphones, one subject at a time. The order of presentation of the tempo conditions was balanced across subjects. Subjects were asked to check on a response sheet whether the sentence they heard contained ratted, rabid or rapid.

RESULTS

Figure 1 gives the combined results of the experiment averaged across all subjects. Looking first at the slow condition (solid line), with a 50 msec gap duration (stimulus 5), we find the subjects heard rabid about 90 percent of the time, while at longer durations they heard increasingly more rapids with the perceptual boundary (50 percent cross-over) occurring at about 75 msec. At closure durations shorter than 50 msec, subjects heard more and more apical flaps with the perceptual boundary occurring at about 35 msec. Stimulus 1, which had the 10 msec closure duration, was the most persuasive stimulus in the series with 100 percent ratted responses in the slow condition.

When the test words were appended to the fast carrier sentence (dashed line), the identification curves shifted systematically toward shorter values. Indeed the fast curve resembles the slow curve compressed in time by about 10 percent. Thus the boundary between /b/ and /p/ shifts about 8 msec. Stimulus 7, for example, which was heard in the slow condition as 64 percent /b/, is now heard as 41 percent /b/. If we look at individual listeners and their responses to stimuli 6 through 12, we find that twelve of the sixteen have fewer /b/ responses for these stimuli in the fast condition than in the slow. This difference was significant at $p < 0.01$ by the Wilcoxon matched-pair signed-ranks test.

The boundary between the /b/ and the flap moves only about 3 msec (roughly 10 percent) in the fast condition. Not only is the shift small, but only nine of the sixteen subjects showed this shift by having more /b/ responses for stimuli 1 through 5 in the fast condition than in the slow. Three subjects had the same number of /b/ responses in the two conditions, while four showed the opposite trend. For the thirteen listeners who did show a difference between conditions, the favoring of /b/ responses was significant at the .05 level.

DISCUSSION

Voicing Effect

The results of this experiment have replicated Lisker's (1957) finding that: (1) an intervocalic /b/ will remain perceptually voiced even when all traces of glottal pulsing have been removed from the closure interval, as long as the closure interval is kept sufficiently short, and (2) if the silent closure interval is lengthened, American listeners will hear a phonologically voiceless stop.

Production data have shown that glottal pulsing is frequently not maintained during closure in this context (Lisker, 1957; Port, 1976) and, as might be expected, perception data indicate that the absence of glottal pulsing is not a sufficient cue for a voiceless medial stop. A similar effect is well-known for the production and perception of English voiced stops in other contexts as well. For example, utterance-initial (Lisker and Abramson, 1964) and word-final voiced stops (Raphael, 1971) are most often not fully voiced through the closure and do not require closure voicing to be heard as voiced (Abramson and Lisker, 1965; Raphael, 1972). There are several reasons

FIGURE 1

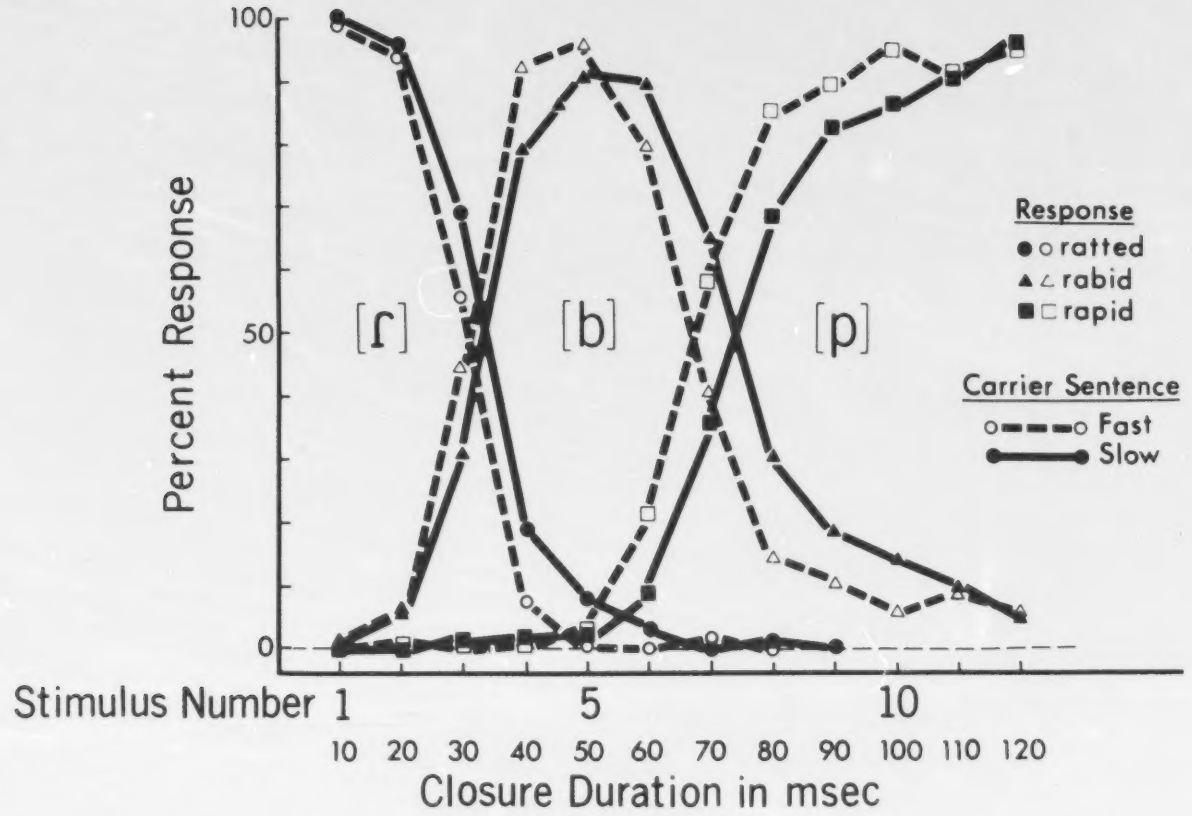

Figure 1: Identification of stimuli as ratted, rabid or rapid plotted as a function of the duration of the silent closure interval between the first and second syllables in both fast and slow carrier sentences.

why glottal pulsing during stop closure might be difficult information for a language to depend on in order to distinguish words. First, glottal pulsing is not easy to maintain during closure because the supraglottal cavities must be enlarged (whether actively or passively) in order to maintain airflow through the glottis (Rothenberg, 1968; Perkell, 1969; Bell-Berti, 1975), and therefore any pulsing that occurs may tend to be rather weak. Second, because the vocal tract is entirely sealed, sound radiation through the tissues is greatly attenuated and therefore likely to be less reliable as a source of perceptual information. Thus a language might not rely entirely on closure voicing to specify a phonological feature in order to distinguish homorganic stops. Indeed Lisker and Abramson (1971) and Abramson (1976) have proposed that this feature in English is manifested primarily in terms of laryngeal timing, with vowel duration and stop closure duration as only secondary phonetic cues for post-vocalic stops. In the present study, it seems the absence of glottal pulsing during the closure interval and the duration of the preceding vowel were ambiguous. Therefore listeners relied on the relative duration of closure to determine the phonological voicing of the stop.

What would have happened if the glottal pulsing information had been different than it was? So far we have only pilot results in which the experimenter was also the subject. These explorations indicate that if audible glottal pulsing is maintained throughout the closure interval, no amount of lengthening will make the stop perceptually voiceless. At extremely long intervals (200-300 msec) it becomes a geminate with a word boundary (rather like raab bid). On the other hand, if the absence of glottal pulsing is maintained for only about 15 msec or more after stop release (that is, if there is any appreciable amount of aspiration), the stop will be heard unequivocably as voiceless for a wide range of closure durations and preceding vowel durations. Thus, there would appear to be a hierarchy of voicing cues. If audible glottal pulsing is maintained from the preceding vowel through to the following vowel, the stop is heard as voiced no matter what the durations of preceding vowel and stop closure. The absence of audible pulsing for a short interval after the stop closure, on the other hand, perhaps because it is an unambiguous indication of vocal fold abduction, induces perception of voicelessness regardless of the vowel and stop closure durations. It is only when the closure interval alone is voiceless that phonological voicing is sufficiently indeterminate for stop closure and vowel durations to be effective cues.

Place of Articulation Effect

The perception of an apical flap for those stimuli with very short artificial closure durations is quite persuasive. The shortest stimulus was almost unanimously identified as an apical flap or a phonological /t/. There is little reason to suspect that masking could account for this effect since both the imploding and exploding transitions could be perceived clearly as labials in isolation. A possible interpretation, however, is that the very short closure durations used here served as an overwhelming and powerful cue for an apical flap that overrode other information in the transition and burst for a labial stop.

To pursue this interpretation, we will first examine production data on the durations of /b/ and apical flaps to see how effective closure duration

might be as a discriminator of the two stops in this context. Port (1976) had speakers read a short carrier sentence containing a medially placed test word with an orthographic representation of one of the six phonological English stops in poststress position. The sentences containing the test words dibber, dipper, didder, ditter, digger and dicker were read at three maximally different speaking tempos. The test words spelled both with tt and dd were uniformly pronounced as flaps by these educated New York city talkers and had identical closure duration distributions at all three speaking tempos. Here we are interested only in data on the fastest tempo speakers could produce. Figure 2 is a frequency histogram of the stop closure durations for medial /b/ and /d/ flaps (intended /t/, as indicated orthographically, was left out in order to retain equal Ns) at the five speakers' fastest tempo. Of the 40 /b/'s at this tempo, six fall into the range between 25 and 35 msec, although none were shorter than 25 msec. The apical flaps, however, ranged as short as 5 msec and as long as about 40 msec. Thus it seems that in this sentence context, even when speakers are urged to speak as quickly as possible, /b/ is rarely shorter than 30 msec, while apical flaps are often shorter. This difference in characteristic timing might serve as perceptual information.

In order to see if a rather different phonetic task might produce shorter /b/ closure durations than these, an auxiliary experiment (not reported elsewhere) was conducted. Nine speakers were encouraged to repeat the nonsense word dababba ([dəbæbə]) in clusters of four or five repetitions on a single intonation contour as rapidly as possible. After some amount of practice (and amusement), a recording was made of several clusters for each subject. The fastest cluster containing true stops was chosen to be analyzed for each speaker, and the durations of the prestress and poststress /b/ closures were measured to the nearest 5 msec on a spectrogram. The mean closure duration for the 35 poststress /b/'s was 45 msec. Only one of these stops was measured as 25 msec and only two were 30 msec in duration. (For the prestressed /b/ the durations were longer; the mean was 65 msec and no stops had closure durations shorter than 50 msec.) The results of this simple experiment indicate that even a simply rhythmical phonetic task does not permit speakers to make lip closures shorter than 30 msec very often. It is plausible to assume, then, that this value is close to a physiological minimum closure duration for a /b/. To the extent that listeners are familiar with these durational characteristics of the production of apical and labial stops, a very brief closure duration could potentially serve as perceptual information for a gesture of apical closure rather than labial closure.

Although it might be tempting to suppose that this coincidence of the perceptual boundary and the production boundary is all that is necessary to account for the effect of flap perception described here, it is surely essential that the formant transitions on the /b/ and flap in vocalic context be minimally distinctive. For example, in pilot exploration it was found that artificial shortening of the /g/ closure in ragged, as pronounced by the same speaker, yielded perception of only a more lax /g/ or, more narrowly, the voiced velar fricative [ɣ]. Even with a closure duration of zero, the word was very unlikely to be confused with ratted, although /g/ is also rarely shorter than about 35 msec (Port, 1976). Examination of spectrograms of the three words rabid, ratted and ragged by the same speaker in Figure 3 suggests a possible reason for this apparent difference between medial /b/ and /g/ in susceptibility to the flap effect in this vowel context. The first formant

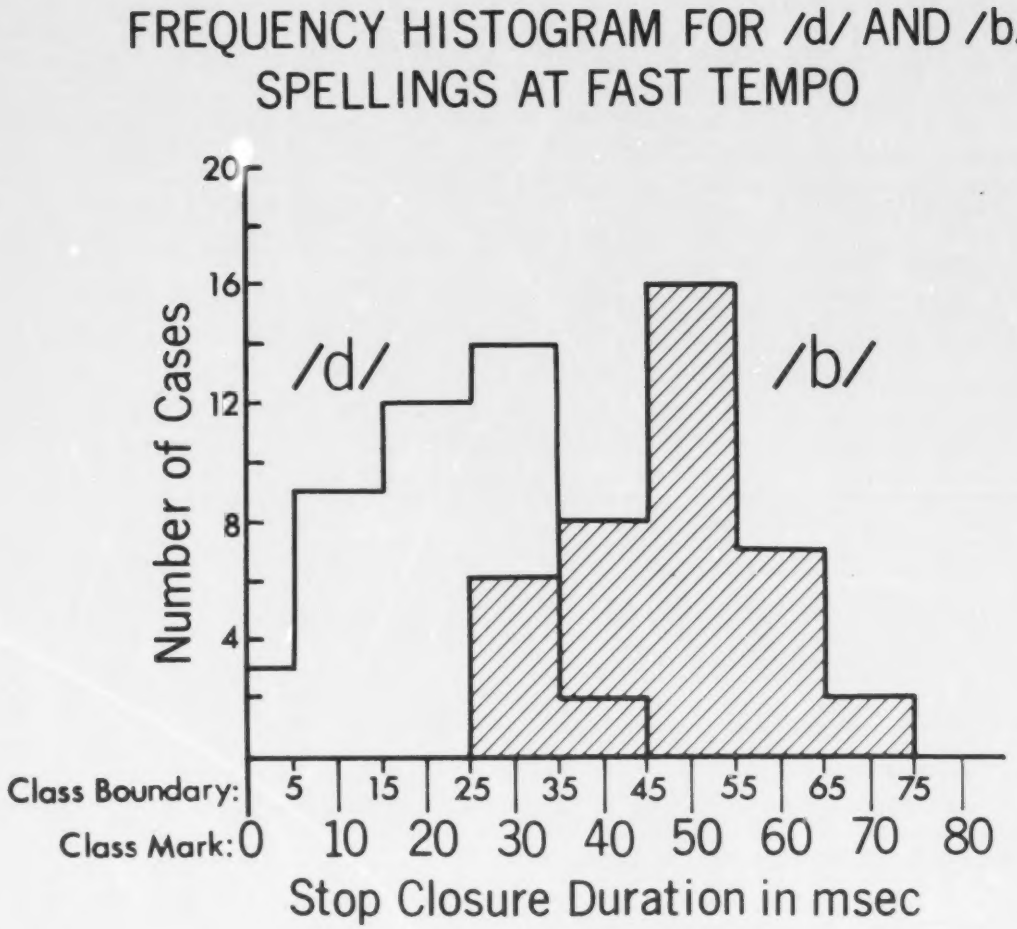

Figure 2: A frequency histogram of stop closure durations for medial post-stress /b/ and apical flaps (for stimuli spelled dd) in carrier sentences spoken at five speakers' fastest tempo. Data from Port (1976).

Figure 3: Sound spectrograms of the words rabid, ratted and ragged spoken in isolation by a male talker.

trajectory is almost identical in the three spectrograms. We will discuss F_2 and F_3 since they are known to contain most information about place of articulation. Looking first at rabid (not the same utterance our perceptual stimuli were made from, but by the same speaker), we find a gentle fall in both F_2 and F_3 approaching the medial stop and a slight rise leaving the stop in the second syllable. For this utterance of ratted there is also a slight dip in F_2 approaching and leaving the medial stop, but with a far briefer duration, while for F_3 there is no apparent transition on either side of the stop. In both these utterances, F_2 and F_3 remain quite parallel after the /r/ constriction. The most obvious difference between these two spectrograms is the duration of the stop closure. On the other hand, for ragged (pronounced with a phonetically identical stressed vowel) we find that F_2 and F_3 display a distinctive converging pattern during much of the syllable preceding the /g/ and then diverge after the consonantal closure. Thus, in the context of these vowels and in this stress pattern, labial and apical stops may be rather weakly distinguished in their pattern of formant trajectories and burst. When the closure duration of a labial stop in such a context is shorter than a labial stop can ever be, listeners are quite persuaded that an apical gesture was performed and thus hear a phonological /t/ or /d/. The velar stop, however, with its more distinctive formant transitions could not be overridden by information from the duration of the closure interval.

In conclusion then, although this effect of shifting place of articulation by means of closure duration alone could probably not be widely generalized from this context, the effect does indicate that a weakness of spectral information for place of articulation may show up even in natural speech in particular contexts (cf. Cole and Scott, 1974; Dorman, Studdert-Kennedy and Raphael, 1977). The perception of a persuasive apical flap for a shortened /b/ closure required both a weak difference in formant transitions and a highly distinctive difference in characteristic closure duration.

The Tempo Effect

When the test words were preceded by the fast carrier sentence, the identification function shifted about 10 percent toward shorter values. Thus the boundary between /b/ and /p/ became about 8 msec shorter in the pooled data, while the /b/ flap boundary moved much less, if at all. These results indicate that listeners based their perceptual criteria for voicing judgments on both the duration of stop closure and the speaker's tempo. The direction of change in the function is in accord with the prediction derived from earlier perceptual results (for example, Pickett and Decker, 1960; Summerfield and Haggard, 1972) and the production data of Port (1976).

Given the presence of a tempo effect on at least the voicing boundary, the question arises as to why the effect is only 10 percent. The fast test sentences were about 30 percent shorter than the slow sentences at midcontinuum (and a few percentage points higher or lower at each end of the set of test words). One might expect a shift in boundary by a comparable percentage (Joos, 1948). There are several likely reasons why the boundary shift is proportionally smaller than overall tempo change. First, the test words received the main sentence stress in these stimuli, and production data (Peterson and Lehiste, 1965; Port, 1976) indicate that speakers change the duration of the sentence-stressed word less than they change other parts of

the sentence when they alter speaking tempo. Thus the stressed word normally occupies a proportionally larger part of the sentence duration at faster tempos. Indeed, since the test words were not altered between tempo conditions, the rab and bid syllables themselves may have provided information for just such durational constancy and thus contributed to a reduction in the tempo effect. A second possible factor is that the target word was placed at the end of the sentence where sentence-final lengthening might also reduce the tempo effect if listeners assume greater sentence-final lengthening is taking place in the fast sentences than in the slow (Oller, 1973; Klatt, 1976). Since our test words were prepared from a slow production of rabid, the stressed vowel and word-final /d/ closure should have been a little too long for the fast carrier sentence, thereby suggesting greater ritenuto on the final word and attenuating the tempo effect. A simple hypothesis incorporating both these factors is the possibility that the voicing cue in this circumstance is determined primarily by the ratio of the stop closure duration to the preceding vowel duration. Since the vowel duration was constant in these stimuli, this might account for the relatively small shift in voicing boundary.

Although it is clear from these findings and others' results that listeners compensate for differences in tempo when evaluating the phonetic significance of durational cues, there remains the important question as to how this compensation takes place. At one extreme, listeners might conceivably extract absolute temporal information (as though in milliseconds) and then perform a "normalization" operation similar to that proposed for spectral information (Joos, 1948; Gerstman, 1968). Alternatively, phonetic timing information might be specified in the nervous system in more abstract terms, such as appropriate durational ratios, that would remain invariant under changes in speaking tempo. These abstract timing relationships could thus be detected directly from the acoustic signal (Martin, 1972; Shankweiler, Strange and Verbrugge, 1977; Turvey, 1977). It is clear from these results, however, that when the closure duration of an intervocalic stop is a major cue for voicing, its phonetic value is influenced by the tempo of its context. It is also possible that this effect interacts with other variables such as the position of the word in the sentence and its stress.

SUMMARY AND CONCLUSIONS

In this study the interaction of two timing variables was investigated for speech perception in English. The duration of the stop closure interval in a natural production of the word rabid was made silent and varied in duration in small steps. Listeners made forced choice identifications of words from the test continuum when appended to carrier sentences spoken at two tempos. The results replicate Lisker's (1957) report that naturally produced medial stops with silent closure can be heard as either voiced or voiceless, depending on the closure duration. These results further demonstrate that the perceptual boundary for stop voicing along such a continuum is sensitive to changes in the tempo of a preceding carrier sentence.

It was found that if the duration of /b/ closure in this utterance of the word rabid was made shorter than about 30 msec, listeners heard a persuasive apical flap. Examination of spectrograms suggested that the formant trajectories of rabid and ratted are minimally different, while their closure

durations were quite distinctive. Some evidence was found that labial stop closure durations are not normally shorter than about 30 msec. Thus it seems that when the closure duration was long enough so that either the lips or tongue tip could have produced the stop, the spectral information dominated and listeners heard a labial stop. But when the closure was shorter than a possible labial stop, the durational cue became unambiguous and was apparently not contradicted by the spectral information. Since both the rab and bid syllables clearly contained labial stops when heard in isolation, this result could not be accounted for in terms of an hypothesis of backward masking.

It is notable that all the perceptual timing effects examined here--voicing, place of articulation and tempo--parallel the timing patterns of speech production whether due primarily to phonological factors (such as stop voicing) or due to physiological constraints (such as the place-of-articulation effect). Speech perception must be understood in terms of a phonetic space whose timing parameters are similar to those of the temporal structure of speech production. Apparently this temporal dimension must be less abstracted from continuous time than segments or digits, since otherwise details of relative timing should not be so essential to the specification of words in both production and perception (Lisker, 1974; Klatt, 1976). On the other hand, phonetic time must be considerably more abstract than absolute intervals, such as msec, since there are strong effects of language-external features such as speaking tempo that modify the timing of speech production and yet are somehow compensated for in speech perception.

REFERENCES

Abbs, M. H. (1971) A study of cues for the identification of voiced stop consonants in intervocalic position. Unpublished Ph.D. dissertation, University of Wisconsin.

Abramson, A. S. (1976) Laryngeal timing in consonant distinctions. Haskins Laboratories Status Report on Speech Research SR-47, 105-112. [Also Phonetica, in press]

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. Proceedings of the 5th International Congress on Acoustics, A51. (Liege: G. Thone).

Bell-Berti, F. (1975) Control of pharyngeal cavity size for English voiced and voiceless stops. J. Acoust. Soc. Am. 57, 456-461.

Cole, R. A. and B. Scott. (1974) Toward a theory of speech perception. Psychol. Rev. 81, 348-374.

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. Haskins Laboratories Status Report on Speech Research SR-17/18, 17-21.

Dorman, M. F., L. Raphael, A. M. Liberman and B. Repp. (1975) Some masking-like phenomena in speech perception. Haskins Laboratories Status Report on Speech Research SR-42/43, 265-276.

Dorman, M. F., M. Studdert-Kennedy and L. J. Raphael. (1977) The invariance problem in initial voiced stop consonants: Release bursts and formant transitions as functionally equivalent context-dependent cues. Percept. Psychophys. 22, 109-122.

Fujimura, O. (1975) A look into the effects of context - some articulatory and perceptual findings. Paper presented at the Eighth International Congress of Phonetic Sciences, Leeds, England.

Gaitenby, J. (1965) The elastic word. Haskins Laboratories Status Report on Speech Research SR-2, 3.1-3.12.

Gay, T., T. Ushijima, H. Hirose and F. S. Cooper. (1974) Effect of speaking rate on labial consonant-vowel articulation. J. Phonetics 2, 47-63.

Gerstman, L. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Joos, M. (1948) Acoustic phonetics. Linguistic Society of America Language Monograph No. 23. (Baltimore: Waverly Press).

Klatt, D. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J. Acoust. Soc. Am. 59, 1208-1221.

Kozhevnikov, V. and L. Chistovich. (1965) Speech: Articulation and Language. Trans. by Clearinghouse for Federal Technical and Scientific Information. (Washington, D.C.).

Lisker, L. (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. Language 33, 42-49.

Lisker L. (1974) On time and timing in speech. In Current Trends in Linguistics vol. 12, ed. by T. Sebeok, et al. (Mouton: The Hague), pp. 2381-2418.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English. Lang. Speech 10, 1-28.

Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 767-785.

Martin, J. (1972) Rhythmic (hierarchical) versus serial structure in speech and other behavior. Psychol. Rev. 79, 487-509.

Massaro, D. (1972) Preperceptual images, processing time and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

Massaro, D. (1974) Preperceptual images, processing time and perceptual units in speech perception. In Understanding Language: An Information-Processing Analysis of Speech Perception, Reading and Psycholinguistics, ed. by D. Massaro. (New York: Academic Press), pp. 125-150.

Oller, O. K. (1973) The effect of position in utterance on speech segment duration in English. J. Acoust. Soc. Am. 54, 1235-1247.

Perkell, J. S. (1969) Physiology of Speech Production. (Cambridge: M.I.T. Press).

Peterson, G. and I. Lehiste. (1965) Duration of syllabic nuclei in English. J. Acoust. Soc. Am. 32, 693-703.

Pickett, J. M. and L. Decker. (1960) Time factors in perception of a double consonant. Lang. Speech 3, 11-17.

Port, R. (1976) The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words. Ph.D. dissertation, University of Connecticut. (Available from the Indiana University Linguistics Club, Bloomington, Indiana 47401.)

Raphael, L. (1971) Preceding vowel duration as a cue to the perception of voicing of American English consonants in word final position. Ph.D. dissertation, City University of New York. (Issued as a supplement to Haskins Laboratories Status Report on Speech Research.)

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of voicing of word-final consonants in American English. J. Acoust. Soc. Am. 51, 1296-1303.

Repp, B. (1976) Perception of implosive transitions in VCV utterances. Haskins Laboratories Status Report on Speech Research SR-48, 209-233.

Rothenberg, M. (1968) The breath-stream dynamics of simple-released-plosive production. Bibliotecha Phonetica No. 6. (Basel: Karger).

Shankweiler, D., W. Strange and R. Verbrugge. (1977) Speech and the problem of perceptual constancy. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. E. Shaw and J. D. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Associates).

Sharf, D. (1962) Duration of post-stress inter-vocalic stops and preceding vowels. Lang. Speech 5, 26-30.

Sharf, D. J. and T. Hemeyer. (1972) Identification of place of consonant articulation from vowel formant transitions. J. Acoust. Soc. Am. 51, 652-658.

Summerfield, A. Q. (1974a) Towards a more detailed model for the perception of voicing contrasts. Speech Perception 3, (Progress Report, Department of Psychology, The Queen's University of Belfast), 1-26.

Summerfield, A. Q. (1974b) Processing of cues and contexts in the perception of voicing contrasts. In Preprints of the Stockholm Speech Communication Seminar, ed. by G. Fant. (Uppsala: Almqvist and Wiksell), vol. 3, pp. 77-86.

Summerfield, A. Q. (1975) Cues, contexts and complications in the perception of voicing contrasts. In Speech Perception, Series 2. (Progress Report, Department of Psychology, Queen's University of Belfast), no. 4, 99-129.

Summerfield, A. Q. and M. P. Haggard. (1972) Speech rate effects in the perception of voicing. In Speech Synthesis and Perception. (Progress Report, Psychology Laboratory, University of Cambridge), no. 6, 1-12.

Summerfield, A. Q. and M. P. Haggard. (1973) Vocal tract normalization as demonstrated by reaction times. In Speech Perception, Report on Speech Research in Progress, Series 2. (Progress Report, Department of Psychology, The Queen's University of Belfast), no. 3, 1-26.

Turvey, M. T. (1977) Contrasting orientations to the theory of visual information processing. Psychol. Rev. 84, 67-88.

Reading Reversals and Developmental Dyslexia: A Further Study

F. William Fischer[†], Isabelle Y. Liberman[††] and Donald Shankweiler[††]

ABSTRACT

The pattern of errors in reading isolated words was studied in two groups of children with respect, particularly, to reversals of letter sequence and letter orientation. One group (the Institute group) consisted of children 8 to 10 years old who had been diagnosed "dyslexic" according to medical and psychoeducational criteria. The other (the School group) included all the children in a second-year elementary school class (see Liberman, Shankweiler, Orlando, Harris and Berti, 1971) who fell into the lowest third on a standard test of reading achievement. Although the Institute children were somewhat poorer in word recognition than the backward readers selected purely on psychometric grounds, the groups did not differ significantly in the incidence of reversal errors. Also, for both groups, reversals represented a small proportion of the total number of reading errors. The performance of the two groups differed in two respects: in relation to directional bias in letter reversals and in the presence or absence of a significant correlation between letter-reversing and word-reversing tendencies. It was concluded from this that directional problems do not loom large in importance in most cases of reading backwardness, but may provide an additional source of difficulty for some dyslexic children. Other aspects of the error pattern were remarkably the same for both groups. The bulk of reading errors made by both groups reflect their common difficulties in phonemic segmentation of words in the lexicon, in phonetic recoding, and in mastery of the orthography--difficulties, in short, with linguistic characteristics of words rather than with their properties as visual patterns.

[†]University of Connecticut, Storrs.

[††]Also University of Connecticut, Storrs.

Acknowledgment: We wish to thank Dr. John Guthrie, formerly associated with the John F. Kennedy Institute, Baltimore, Md., for his kindness in permitting us to study the children (here referred to as the Institute group) and for making the findings of the Kennedy staff available to us. This work was supported in part by a grant from the Research Foundation of the University of Connecticut, and in part, by a program project grant to Haskins Laboratories from NICHD.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

INTRODUCTION

Use of the label "dyslexic" can be said to express an intent to be more specific than might be implied by use of the designation, "backward in reading." At the very least, the term is ordinarily reserved for those whose achievements in reading are markedly below what would be reasonably predicted from other information we have about the child: his intelligence level, his social-cultural background and his achievement in other academic areas. A dyslexic child, in short, might be distinguished from others who are backward in reading by the presence of a significant disparity between expectation and performance (Benton, 1975).

This disparity is certainly a necessary condition for use of the term. But, of course, more has usually been implied. The designation "dyslexic" carries an implication about the cause of reading failure; it assumes an underlying constitutional inadequacy in one or more of the abilities requisite for reading. It has been supposed, therefore, that there should be signs by which a dyslexic child can be recognized. The medical people who first described the occurrence of reading disability in normally bright children sought deficits in basic perceptual and cognitive functions characteristic of those exhibited by adults with damage or disease of the posterior cerebral hemispheres (Hinshelwood, 1917). Although many correlates have been proposed in the long history of clinical study of dyslexia, there are still no generally accepted criteria that can unequivocally be applied to distinguish the dyslexic child from others who are backward in reading.

It can be argued that one possible reason for failure to define dyslexia satisfactorily stems from insufficient attention to the reading process itself (Shankweiler and Liberman, 1972). If the underlying deficits, which have proved so elusive, are indeed specific to reading and reading-related tasks, then we would be well advised to look very closely at the kinds of reading errors so-called dyslexic children make. One classification of error to which diagnostic significance has been attributed is the tendency to read letter sequences in reversed order and to reverse the orientation of individual letters. Although it has been proposed that dyslexic children are especially prone to reversal tendency (Orton, 1925, 1937), the belief has never been put to a systematic test.

The question could not meaningfully be raised without data on the kinds of errors normal children make at various stages of learning to read. An earlier study (Liberman, Shankweiler, Orlando, Harris and Berti, 1971) explored the occurrence of reversal errors in an entire school population of second graders in an elementary school. It was found that letter confusions and reversals of sequence occurred with appreciable frequency only among the children in the lowest third of the class on a standard test of reading achievement. Even among those children, reversals of order and sequence accounted for only 10 and 15 percent, respectively, of the total of misread letters. Moreover, within this group of poorer readers, only some reversed to an appreciable extent. Thus, poor readers who are slightly beyond the earliest stage of reading acquisition are not generally characterized by a high rate of reversal errors. This raises the possibility, which we consider in the present study, that reversal errors, though not characteristic of poor readers in general, might serve to distinguish reading disability of a

specific kind, from those backward in reading from diverse causes.

To investigate this possibility, we focused on a special population of severely retarded readers selected by the staff of the John F. Kennedy Institute, Baltimore, on the basis of both psychoeducational and medical criteria. The Institute group, chosen as it was to meet conventional criteria for the diagnosis of dyslexia, might be expected to differ in various ways from the group of poor readers studied by Liberman et al. (1971) that had been selected by means of IQ and standard reading test scores alone. Our purpose in the present study was to determine the incidence of commission of reversal errors in the "dyslexic" group with the same test materials from which we had obtained our earlier findings on second grade school children from a Connecticut town. We hoped, thereby, to discover whether children designated as dyslexic exhibit a distinctive pattern of reading errors, and, in particular, to discover whether the pattern of their reversal errors in letters and words differs from that of other poor readers.

MATERIAL AND METHOD

Subjects

The subjects for this study were drawn from a group of children who, because of their extreme reading disability, had been selected for inclusion in a special remedial program at the John F. Kennedy Institute in Baltimore, Maryland. This group had been chosen on the basis of an extensive series of multidisciplinary evaluations carried out by members of the medical and psychological staff of the Johns Hopkins Hospital. The children selected by the Institute were retarded in reading by at least 18 months according to test norms (Gray Oral Reading Test, Form A) based on age. Only those children in good health and with average or above-average intelligence were included. Extensive neurological and psychological screening excluded children with gross signs of brain damage, end-organ deficiencies or severe psychiatric problems. Those children exhibiting the soft signs of neurological dysfunction (for example, general awkwardness, mixed and/or confused laterality) with the exception of strabismus, were allowed to remain in the study sample.

In order to replicate insofar as possible the age and sex characteristics of the subjects used in the Liberman et al. (1971) study, the subjects for the present investigation were further limited to boys between the ages of 8 and 10 (mean age = 8.9 years). The total number meeting these criteria in the Institute group was 13. The children ranged from the second to the fourth year of elementary school.

Procedure

The children were investigated using the same tasks and procedures as those employed by Liberman et al. (1971). The tasks were given individually to all subjects on successive days. A Word List, which is described below, was administered twice to each child, the list order being reversed on the second day. Data from the two presentations of the list were combined in scoring the responses of each subject, but were available separately for assessment of test-retest reliability.

1. Word list. The list comprised 60 monosyllabic words including a selection of primer-level "sight" words, most of the commonly-cited reversible words and, in addition, a group of consonant-vowel-consonant (CVC) words that provide ample opportunity for reversing letter orientation. Each word was hand-printed in manuscript form on a separate 3" x 5" white index card. The children were asked to read each word aloud to the best of their ability.

2. Recognition of briefly-exposed single letters. The test comprised 100 items in which a given letter was to be matched to one of a group of five, including four reversible letters in manuscript form[1] (b, d, p, g) and one nonreversible letter (e) that was added as a reliability check. There were 20 such items for each letter. The order of the resultant 100 items was randomized, as was the order of the multiple choice sequence for each item on the answer sheet. The standard was presented tachistoscopically for matching with one of the multiple-choice items on the answer sheets. Tachistoscopic exposure of the 2" X 2" slides of each letter was projected for 1/125 sec in the center of an 11" X 14" screen mounted 5 feet in front of the subject. A brief training session preceded the presentation of the test stimuli.

Error Analysis of Word Transcription

The children's responses to the word list were recorded on magnetic tape and also phonetically transcribed by the examiner. Scoring procedures were the same as those established by Liberman et al. (1971). Five categories were included in the scoring.

1. Reversals of sequence (RS). Scored when a word or a part of a word was read from right to left (for example, when lap was read as [pæl] or [pleI]; form as [frʌm]).

2. Reversals of orientation (RO). Scored when b, d, p and g were confused with each other, as when bad was read as [dæd], [pæd] or [bæg]. If bad was given as [dæb], it was scored as a sequence error instead. Both types of reversal were scored when nip was read as [bIn].

3. Other-consonant error (OC). Included all consonant omissions and additions as well as all consonant substitutions other than reversals of orientation. A response could contain both a sequence reversal and a consonant error, as in the case of the response [træp] for the stimulus word pat. It could also contain both an orientation reversal and a consonant error, as in the case of the response [træp] for the stimulus word tab. However, confusions among b, d and g were scored only as reversals of orientation, not as consonant errors.

---

[1]The letter g, is, of course, a distinctive shape in all type styles, but it was included among the reversible letters because, historically, it has been treated as a reversible letter. It indeed becomes reversible when printed with a straight segment below the line. (In manuscript printing, as was used in preparing the stimulus materials for this study, the tail of the g is the only distinguishing characteristic.)

4. Vowel error (V). Included all vowel substitutions, such as [pIg] for peg. A vowel error was not charged when a consonant error in the response forced a change in the pronunciation of the vowel, provided the vowel sound produced in the response was a legitimate pronunciation of the original printed vowel (for example, response [ræt] for the stimulus word raw).

5. Total error (TE). Simply the sum of all preceding error types.

RESULTS

The Institute children, diagnosed "dyslexic" after extensive clinical assessment (that included IQ and school achievement testing among its components), are to be compared with a group of elementary school children selected by Liberman et al. (1971) purely on the basis of standard tests of IQ and school achievement. The latter group consisted of all of the children from a public school second grade who met the criteria of testing average or above in IQ and fell in the lower third of the grade in reading achievement. The two groups are fairly well matched for IQ, since both selection procedures excluded children of low IQ. For the Institute children, the Verbal Scale quotient on the WISC ranged from 90 to 140 with a mean VIQ of 107. Liberman et al. (1971) reported for their group a Full Scale IQ range from 85 to 126 with a mean IQ of 99.

Severity of the Reading Deficit

It is of initial interest to discover whether the groups of poor readers selected by quite different criteria do actually differ in the degree of backwardness in reading.

On the Gray Oral Reading Test, which measures the reading of connected prose, the difference between the Institute children and those studied by Liberman et al. (1971) was minor. The Institute group achieved a mean level of performance equivalent to 1.4 years of schooling, while the Liberman et al. (1971) children earned a mean reading score of 1.7 years.

However, a task requiring the children to decode isolated words differentiated the groups more sharply than the Gray test. On this analytic reading test (the "Word List"), the Institute group encountered more obvious difficulty, making significantly more errors ($p < .05$). Thus, while both groups of children were lacking in decoding skills, the Institute group was somewhat more deficient in these skills and had learned to recognize significantly fewer words.

Reversals in Relation to Other Errors

Mean errors per subject in the various error categories derived from the Word List are shown in Table 1. In addition, the table gives the data for errors in recognition of singly presented, tachistoscopically exposed reversible letters. Since the opportunities for error among the different error types were not constant, the data were analyzed by basing percentages on opportunities for error. This analysis reveals that although the Institute children were poorer readers, the distribution of errors among the various error categories paralleled the pattern produced by the Liberman et al. (1971)

TABLE 1: Errors by the Institute and School Groups on the Word List and the Letter Recognition Test, presented as a function of opportunities for error.

	Reversed Sequence		Reversed Orientation		Other Consonant		Vowel		Single Letter Recognition	
	Inst.	Sch.	Inst.	Sch.	Inst.	Sch.	Inst.	Sch.	Inst.	Sch.
Mean Errors	10	8	10	11	50	25	50	33	14	7
Opportunities for Error	120	120	88	88	152	152	124	124	100	100
Percent	8.3	6.7	11.4	12.5	32.9	16.4	40.3	26.6	14.0	7.0

school sample.

In both groups the vowel and other-consonant categories accounted for the bulk of all errors made. The Institute children made more errors in reading other consonants and vowels, demonstrating that they are the less proficient group in reading. However, in spite of their greater reading deficit, it is of major interest to note that the Institute children made relatively the same proportion of reversal errors in reading words as did the children in the school sample (8.3 percent vs. 6.7 percent, respectively, for reversals of sequence; 11.4 percent vs. 12.5 percent for reversals of orientation). Thus, frequency of the two kinds of reversal errors is not an aspect of reading performance that distinguishes these two groups of children.

Although the proportion of the two kinds of reversal errors was the same for the two groups, a discrepancy in their association was noted. Reversals of sequence and reversals of orientation were correlated among the Institute children ($r = .55$; $p < .05$) in reading the word list, but not among the children from the School sample ($r = .03$). These statistics indicate that whereas both groups of children reverse letters and words with roughly the same overall frequency, the children of the Institute group were more consistent in their reversal error, tending to reverse both orientation and letter sequence. For them the situation is as Orton (1937) supposed: the two kinds of reversals are associated. In contrast, the absence of a correlation between sequence and orientation errors among the School sample means that within that group an individual's frequency of reversing letter sequence cannot be predicted from his frequency of reversing letter orientation.

Reversed Orientation of Letters: the Nature of the Confusions

Perception of reversible letters was studied in two ways: by embedding them in words and by presenting them in isolation. A comparison of the frequency and distribution of errors on these different tasks enables us to separate linguistic and contextual contributions to the error rate from the contribution that is purely visual.

In the task of recognition of singly presented reversible letters, the Institute group made appreciably more errors than the School group (14.0 percent vs. 7.4 percent), as can be seen in Table 1. However, examination of the individual subject data in the Institute group reveals that the two subjects with the highest number of errors made at least twice as many errors as the third ranking subject. Omitting the data for these two subjects and recalculating the mean gives 9.3 percent, which is only slightly higher than the figure obtained for the School group. We then see that it is in the main true for the Institute group, as it was for the School group, that more reversals of b, d, p and g occur in the context of reading words than in recognizing these letters in isolation.

Confusions among the four reversible letters in word context are presented in matrix form in Table 2. The matrix shows, for a given letter, the frequency with which it was correctly read or replaced by another phoneme. The column at the left of the matrix lists the reversible letters and each row of the matrix gives the distribution of responses to a given letter made by the children in oral reading. The error frequencies are expressed as

percentages of the total occurrences of each letter in the list (that is, in terms of opportunities for error).

TABLE 2: Confusions among reversible letters in Word List. Percentages based on opportunities.

Presented	Obtained Group*	b	d	p	g
b	Inst.		17.0	6.1	2.2
	Sch.	-	10.2	13.7	0.3
d	Inst.	10.1		1.0	1.4
	Sch.	10.1	-	1.7	0.3
p	Inst.	3.3	0.8		2.6
	Sch.	9.1	0.4	-	0.7
g	Inst.	0	1.3	0.4	
	Sch.	1.3	1.3	1.3	-

*Inst.-Institute group
Sch.-Public school group

Several similarities between the two groups of children may be noted from inspection of Table 2. For both groups, the errors were essentially confined to the truly reversible letters, b, d, and p. The letter b presented the greatest difficulty for both groups, followed by the letters d and p, respectively. In misreading these letters, both groups of children confined their substitutions to letters of similar form. Substitutions other than b, d or p rarely occurred.

A question of considerable interest concerns the directional characteristics of the reversals of letter orientation. The bulk of the misreadings of these letters for both groups involved a substitution from within the set (b,d,p,g). Moreover, most of these errors were produced by a single 180 degree rotation. Thus, rotations that occurred in the horizontal plane produced confusions among b and d, or g and p, and those in the vertical plane produced confusions among b and p, or g and d.

The frequencies of horizontal and vertical transformations of the set are presented in Table 3. Data for the Word List are shown in the top row. It may be seen that the Institute children differed from the School sample in making a disproportionate number of reversal errors in the horizontal plane. The School group made horizontal and vertical reversals with about the same frequency.

TABLE 3: Horizontal and vertical transformation of reversible letters. Percentages based on opportunities.

	Institute Group		School Group	
Task	Horizontal*	Vertical†	Horizontal	Vertical
Word Context	7.5	2.8	5.6	6.1
Single Letter Recognition	9.3	4.0	3.5	1.1

*b-d, d-b, p-g, g-p
†b-p, p-b, d-g, g-d

TABLE 4: Directional scan of horizontal transformations of reversible letters. Percentages based on opportunities.

	Institute Group		School Group	
Task	Left-to-Right	Right-to-Left	Left-to-Right	Right-to-Left
Word Context	5.0	10.0	5.7	5.5
Single Letter Recognition	9.5	9.0	5.0	2.0

The frequencies of horizontal and vertical rotations for multiple-choice recognition of the briefly exposed single letters are given in Row 2 of the table. Here we find for both groups a predominance of rotations in the horizontal plane. On both tasks, the Institute children made errors involving horizontal transformation more than twice as often as errors involving a vertical rotation. The School children, on the other hand, showed a bias (favoring horizontal rotation) only when perceiving the letters out of word context.

We may further examine the directional characteristics of the reversals of orientation by noting any asymmetries in frequency of reversing from left-to-right and from right-to-left. By rotating the axis in a left-to-right direction, one transforms the letter d to b, whereas rotation of the axis from right-to-left transforms b to d. From inspection of Table 4, we discover a definite difference in the behavior of the two groups. Among the Institute children, horizontal transformations involving a rotation in the right-to-left direction occurred twice as often as rotations in the left-to-right direction. Thus, in these children, confusions such as reading b as d occurred considerably more often than errors where d is read either as p or b. The bias of the Institute Children for an excess of right-to-left confusions is present only in reading words, however, not in recognition of individual isolated letters. In the School group, on the other hand, no directional bias was found in reversal errors that occurred in reading words. In recognition of briefly exposed single letters, this group showed a tendency to reverse from left-to-right.

Distribution of Errors Within The Syllable

Notwithstanding these differences in the directional characteristics, the overall error pattern for both groups of children is remarkably the same. This is reflected both in the relative distribution among the various error types and in the position of errors within the syllable. As can be seen in Figure 1, both groups of children, when reading words, made approximately twice as many other-consonant errors in the final segment of the syllable following the vowel, as in the initial segment preceding the vowel. However, this effect of position in the syllable did not occur with reversible consonants. For both groups of children, reversible consonants produced equal numbers of errors in initial and final position. Thus, it would seem that reversible consonants present extra problems that override other difficulties children have in analyzing the structure of the word. At all events, although the error pattern within the syllable differs for reversible and nonreversible consonants, it differs in identical fashion for the "dyslexic" children and for those poor readers identified purely on a psychometric basis.

DISCUSSION

The experiment was designed to compare the reading reversal tendencies of children diagnosed as dyslexic with those of a group of children reading in the lower third of a public school class. Before the question of reversal tendency could be explored, however, it was of some concern to determine whether the Institute (that is, dyslexic) group differed from the School sample in severity of overall reading backwardness. Although there was only a small difference between the mean scores of the two groups on a conventional

FIGURE 1

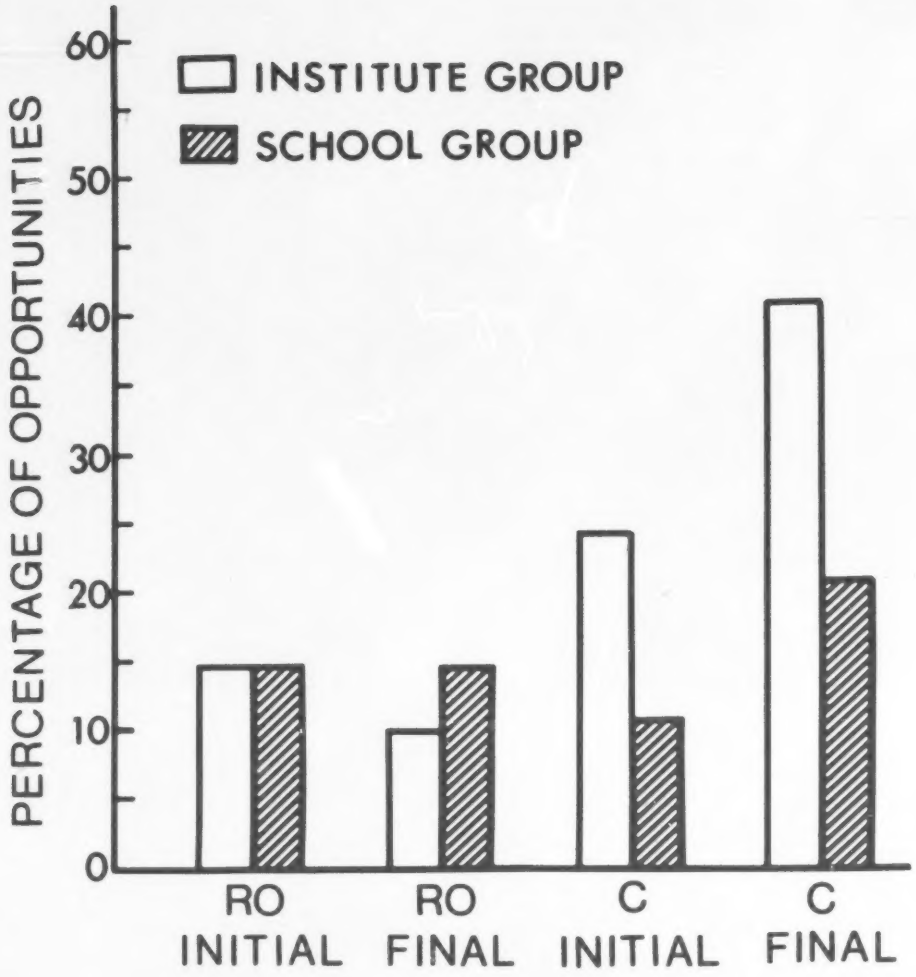

Figure 1: A comparison of errors according to syllable position, percentages based on opportunities for error.

test of reading prose (Gray Oral Reading Test, Form A), more marked differences in reading ability emerged when the children were required to demonstrate their decoding skills without the assistance of contextual cues. The Institute children were unquestionably less proficient in decoding isolated words, making significantly more errors than the School group. We conclude, therefore, that whatever the label "dyslexic" might mean, it is clear that in our sample these children are the poorer readers.

Having established this, it remained to determine whether the two groups of children differed qualitatively in the nature of their reading difficulty. Since the claim is frequently made that reversal errors are more prevalent among dyslexic children, it was of critical interest to determine whether the two groups of children could be distinguished on the basis of reversal tendency.

Results obtained with the School group (Liberman et al., 1971) indicated that by the end of the second year at school, reversals occur in appreciable number only among the poorest readers. The Institute group, though in general lower in reading proficiency than the children in the School group, did not differ significantly from them in incidence of either word or letter reversals. Moreover, in both groups reversal errors accounted for only a small proportion of the total errors in reading the test words. That is, errors made by both groups on other consonants and vowels were far more frequent than reversal errors of either kind.

A further qualitative analysis of the errors on reversible letters was made possible by the fact that they were presented both in isolation and embedded in word context. Comparison of the frequency and distribution of errors on reversible letters under these two conditions enables us to separate linguistic and contextual contributions from the contribution that is purely visual. The results showed that with the exception of two Institute subjects, children in both groups made more letter reversals in the context of reading words than in recognition of the reversible letters in isolation. Thus, with the exception of the two Institute children, it seems that the problem with reversals cannot be attributed primarily to the perception of the individual letter forms.

Although no difference in the incidence of reversal errors was found between the two groups, discrepancies in the directional characteristics of the reversals were observed. In misreading reversible letters, the Institute group made more than twice as many reversals in the horizontal as in the vertical plane. In addition, these horizontal reversals were asymmetric in direction, showing a 2:1 bias toward right-to-left transformation as opposed to left-to-right transformation. The excess of horizontal reversal errors over vertical reversals occurred both in reading words and in recognition of isolated letters, whereas the sinistrad directional bias was specific to the reading of words.

The School group also tended toward horizontal reversals on the isolated letters, but this asymmetry was not found in their reading errors on the Word List. Moreover, no sinistrad directional bias was present in the School group's errors on either the words or the isolated letters. The results for

that group, as noted by Liberman et al. (1971), therefore, did not support Orton's (1937) view that reversals are symptomatic of a tendency to scan words in a sinistrad manner, from right to left.

The Institute children, however, are clearly different. Qualitative analysis of their reversal error pattern, as we have seen, together with the fact that sequence reversals and letter reversals are correlated in this group, consistently point to the failure of these children to establish stable left-to-right habits of scan. Therefore, though Orton's (1925, 1937) emphasis on the reversal tendency as a diagnostic characteristic of the dyslexic reader is not substantiated here, his contention that dyslexics may be distinguished by sinistrad directional bias in their reversals is given some support.

There are indications (Braine, 1968, 1972) that young children have a tendency to begin their inspection of a pattern from the right rather than the left, differing in this regard from older children and adults, who begin from the left. It is conceivable that the Institute group is behaving much as younger normal children do in their tendency to attack words from the right. Zangwill and Blakemore (1972) have noted sinistrad scan in a young adult with a history of dyslexia. These findings clearly warrant renewed developmental study of the problem.

It is also noteworthy that two children from the Institute group differed from the remainder of their group as well as from the School sample in their performance on the task requiring recognition of tachistoscopically presented reversible letters. Whereas the remainder of the children made relatively few errors in recognition of single letters, these children made more errors here than on the word reading task. Their high error rate in recognition of tachistoscopically exposed letters may be related to perceptual factors specific to rapid exposure. It may also be that they are exhibiting a pattern of performance typical of younger children who tend to confuse visual forms that differ in orientation (see Gibson, Gibson, Pick and Osser, 1962). These possibilities should be explored further.

The analysis of reading errors in terms of linguistic categories revealed more similarities than differences in the two groups. Although the Institute children made quantitatively more errors in decoding words, the relative distribution of error types for each group was essentially the same. In both groups the incidence of vowel and other-consonant errors far exceeded the incidence of reversals of sequence and orientation. Moreover, the two groups did not differ significantly in the frequency with which they made reversal and other-consonant errors. One difference that did emerge concerned the vowels: while vowels were the more frequently misread category in both groups, the Institute children made significantly more errors in this category ($p < .05$) than the children from public school.

The preponderance of vowel errors in reading comes as no surprise. A number of earlier studies of beginning and disabled readers (Monroe, 1932; Weber, 1970; Shankweiler and Liberman, 1972) have documented the fact that vowels elicit more errors than consonants. This remains true despite systematic variation of the position of the vowel within the word (Fowler, Liberman and Shankweiler, in press). It has been proposed that the difficulty in decoding vowels may arise from their variable orthographic representation

(Shankweiler and Liberman, 1972; Liberman, 1973), as well as from the continuous nature with which they tend to be perceived (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Liberman, 1970). Consonants, on the other hand, are more strongly categorically perceived and few have multiple orthographic representations. For these reasons, they might, therefore, be expected to elicit fewer errors. Moreover, if vowel representation is indeed more difficult, then one would expect that children with generally inferior decoding skills would have correspondingly more difficulty with vowels.

In addition to examining the error pattern by linguistic category, we also looked at the errors in relation to their position within the syllable. Here the two groups showed identical patterns. In each group, other-consonant errors occurred more frequently in the final segment of the syllable than in the initial segment, whereas errors involving reversible consonants occurred equally often in the initial and final portions of the stimulus word. The position effect for other-consonant errors has been observed in a number of previous studies (Daniels and Diack, 1956; Weber, 1970; Shankweiler and Liberman, 1972; Liberman, 1973). It has been proposed in some of our earlier work (Liberman, 1971; Liberman, Shankweiler, Fischer and Carter, 1974) that in order to read, a child must be consciously aware of the phonemic segmentation of the spoken word and that the position effect may reflect the child's inability to perform that segmentation. The presence of a reversible consonant, the exact identity of which they are unsure, would, of course, be expected to nullify such a position effect for these children.

The persistence of reversal errors may be more important than their frequency of occurrence in identifying the more severe cases of dyslexia or in identifying dyslexia in older children. It is not unreasonable to suppose that older dyslexics may persist in making reversal errors after the age at which these errors normally disappear in other backward readers. Nevertheless, we might also expect that regardless of any differences in reversal tendency to be found in the reading performances of older dyslexics, their overall error pattern and that of other backward readers of normal intelligence will in important ways remain the same. That is, the principal source of reading errors for both groups would continue to be at an altogether different level. The common error pattern, as we have seen, reflects difficulties in phonemic segmentation of words in the lexicon, in phonetic recoding, and in mastery of the orthography--difficulties, in short, with the linguistic characteristics of words rather than with their properties as visual patterns.

REFERENCES

Benton, A. L. (1975) Developmental dyslexia: neurological aspects. In Advances in Neurology, vol. 7, ed. by W. J. Friedlander. (New York: Raven Press).

Braine, L. G. (1968) Asymmetries of pattern perception observed in Israelis. Neuropsycholog. 6, 73-88.

Braine, L. G. (1972) A developmental analysis of the effect of stimulus orientation on recognition. Am. J. Psychol. 85, 157-188.

Daniels, J. C. and Diack, H. (1956) Progress in Reading. (Nottingham: University of Nottingham Institute of Education).

Fischer, F. W. (1972) An analysis of reversal errors in children with severe

reading disability: the relationship to certain linguistic and perceptual factors. Unpublished master's thesis, University of Connecticut.
Fowler, C., Liberman, I. Y. and Shankweiler, D. (in press) On interpreting the error pattern in beginning reading. Lang. Sp.
Gibson, E. J., Gibson, J. J., Pick, A. D. and Osser, R. (1962) A developmental study of the discrimination of letter-like forms. J. Comp. Physiol. Psychol. 55, 897-906.
Gray, W. S. (1967) Gray Oral Reading Test. (New York: Bobbs-Merrill Co.).
Hilgard, E. R. (1962) Methods and procedures in the study of learning. In Handbook of Experimental Psychology, ed. by S. S. Stevens. (New York: John Wiley and Sons).
Hinshelwood, J. (1917) Congenital Word-blindness. (London: H. K. Lewis).
Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.
Liberman, A. M., Cooper, F. S., Shankweiler, D. and Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Liberman, I. Y. (1973) Segmentation of the spoken word and reading acquisition. Bull. Orton Soc. 23, 65-77.
Liberman, I. Y. (1971) Basic research in speech and lateralization of language: some implications for reading disability. Bull. Orton Soc. 21, 71-87.
Liberman, I. Y., Shankweiler, D., Fischer, F. W. and Carter, B. (1974) Explicit syllable and phoneme segmentation in the young child. J. Exp. Child Psychol. 18, 201-212.
Liberman, I. Y., Shankweiler, D., Orlando, C., Harris, K. S. and Berti, F. B. (1971) Letter confusions and reversals of sequence in the beginning reader: implications for Orton's theory of developmental dyslexia. Cortex 7, 127-142.
Monroe, M. (1932) Children Who Cannot Read. (Chicago: University of Chicago Press).
Orton, S. T. (1925) Word-blindness in school children. Arch. Neurol. Psychiat. 14, 581-615.
Orton, S. T. (1937) Reading, Writing and Speech Problems in Children. (New York: W. W. Norton).
Shankweiler, D. and Liberman, I. Y. (1972) Misreading: a search for causes. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).
Weber, R. (1970) A linguistic analysis of first-grade errors: A survey of the literature. Read. Res. Quart. 4, 96-119.
Zangwill, O. L. and Blakemore, C. (1972) Dyslexia: reversals of eye-movements during reading. Neuropsychol. 10, 371-373.

The Noncategorical Perception of Tone Categories in Thai*

Arthur S. Abramson†

ABSTRACT

Arguments continue over categorical perception of phonetic segments versus continuous perception. In tone languages, phonemic tones are characterized principally by F_0 contours. Theories of categoricity would predict continuous perception of tones; that is, there ought to be no peaks in the discrimination of variants at category boundaries, more as in experiments with isolated synthetic vowels than with synthetic stop consonants. This was the outcome of earlier work (Abramson, 1961), but it seems to have been contradicted by recent work on Mandarin (Chan, Chuang and Wang, 1975). The present study involves a considerably larger number of subjects, 34 native speakers of Thai--a language with five phonemic tones. Sixteen flat F_0 variants synthesized on a syllable of the type [kha:] were sorted into the three "static" high, mid and low tones with considerable overlap. Discrimination tests yielded a high level of discrimination across the continuum with no effects of boundaries between categories, thus implying noncategorical perception of tone categories.

BACKGROUND

Theories of the categorical perception of speech seek to explain how levels of acuity of discrimination vary with the type of phonetic segment involved. The now classical paradigm requires the preparation of a continuum of variants along some physical dimension susceptible to auditory division into phoneme categories. Zones of ambiguity between the groups of variants labeled as phonemes are the so-called phoneme boundaries. In the ideal case of categorical perception, subjects will not do much better than chance in

---

*This is a slightly revised version of an oral paper given before the 93rd Meeting of the Acoustical Society of America, The Pennsylvania State University, State College, Pennsylvania, 6-10 June 1977.

†Also The University of Connecticut, Storrs.

Acknowledgment: This research was supported by a grant to Haskins Laboratories from the National Institute of Child Health and Human Development. Although the stimuli for this research were made at Haskins Laboratories and the data were analyzed there, the data themselves were gathered while the author was on sabbatical leave in Thailand on research fellowships from the American Council of Learned Societies and the Ford Foundation Southeast Asia Fellowship Program.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

discriminating variants within the labeled categories, but will show rather high peaks of discrimination in the regions of the phoneme boundaries. This is best shown in experiments with stop consonants (Liberman, Harris, Hoffman and Griffith, 1957). A rather different set of results prevails with steady-state isolated vowels (Fry, Abramson, Eimas and Liberman, 1962). Variants are indeed grouped into phoneme categories but with somewhat more overlap than for stops; however, discrimination is very good along the whole continuum without any special effects at the category boundaries. Two ways of explaining these differences have dominated the literature. The motor or articulatory-reference theory (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967) points out that stop consonants and presumably certain other types of phonetic segments are essentially discontinuous in their mode of production, while vowels are continuously graded in production. Thus, it is claimed that the extent of categoricity in perception is shaped at some psychological level by the intervention of these radical differences in production. The other theory invokes short-term memory. Consonants, particularly stop sounds, have rapidly changing spectra, while vowels, certainly steady-state vowels, tend to be rather long in duration providing for more time to process the stimuli with less strain on memory (Fujisaki and Kawashima, 1969).

Arguments continue over the categorical perception of phonetic segments versus continuous perception. There is still a need for experiments on types of phonetic segments not previously examined and types of segments that have been insufficiently examined. For example, experiments with voice onset time (Abramson and Lisker, 1970) yield data very similar in categoricity to those of data with stop consonants differing in place of articulation, while experiments with distinctive vowel length in Thai (Bastian and Abramson, 1962) yield data very similar to those for steady-state vowels. In 1961, I presented before this society some work on the identification and discrimination of phonemic tones in Thai (Abramson, 1961). A set of five fundamental-frequency contours with final points moving upward incrementally from a level base, all with the same short drop at the end, was divided perceptually by native speakers into the mid and high tones. In ABX tests, the overall level of discrimination was very high with no convincing evidence of a discrimination peak at the category boundary, thus confirming a prediction of continuous perception. On the other hand, in a recent paper before this society, Chan, Chuang and Wang (1975) presented contradictory results showing boundary effects for a fundamental-frequency continuum that yielded two of the four tones of Mandarin Chinese, namely the rising and level tones. The present study is an attempt to explore the matter further.

EXPERIMENT

Central Thai or Siamese has five lexical tones characterized for the most part by different fundamental-frequency contours (Abramson, 1962; Erickson, 1976). They are conventionally divided into static tones--high, mid and low--and dynamic tones--falling and rising. There appears to be no single acoustic dimension along which all five of these tones lie. In my desire to run an experiment analogous to the study of the English stop consonants /bdg/, I sought a continuum that would yield three tonal categories, namely the low, mid and high tones. In addition, since I had a large number of subjects available to me, I wished to have a continuum with varying degrees of acceptance on the part of native speakers of Thai for division into those

three tones. I thought that I might thus have a better chance to test the perceptual effects of phoneme categories. That is, if perception of tones is categorical rather than continuous, those subjects who showed good solid identification functions might also show high peaks of discrimination at the category boundaries, in contrast with those who did not readily sort the stimuli into tones. I did this by ignoring the small but noticeable movements of these "static" tones as normally produced and by using level fundamental frequencies instead.

The Haskins Laboratories parallel-resonance synthesizer was used to produce the stimuli. The basic pattern was a set of steady-state formant frequencies chosen to yield a vowel acceptable as Thai long /aa/ with initial formant transitions appropriate to the velar place of articulation. Voiceless aspiration was simulated by providing a voicing lag of 80 msec filled with turbulent noise in the regions of the upper formants, with the first formant absent during the lag. The voice source was turned on at the end of the voicing lag, and one of the set of fundamental-frequency levels was imposed. The overall amplitude was kept flat throughout the syllable except for a slight rise at the beginning and a slight fall at the end. These specifications yielded syllables of the type [kha:]. Sixteen fundamental-frequency levels ranging from 92 Hz to 152 Hz in steps of 4 Hz were imposed on the basic pattern. These 16 synthetic syllables were presented in a number of random orders to 33 native speakers of Thai for identification as one of five possible words. In fact, it was expected that only three of the choices would be used: /khàa/ (low) 'galangal, a rhizome', /khaa/ (mid) 'a grass' and /khá̀a/ (high) 'to engage in trade'.[1]

Of the various discrimination procedures used in speech research, I chose the "four-interval forced-choice test of pair similarity" (4IAX) which, according to Pisoni (1971), has certain advantages over other methods. For each trial the subject is presented with two pairs of stimuli such that one pair contains identical members and the other pair contains members that differ along the stimulus dimension. All possible 4IAX arrangements to provide one-step and two-step comparisons along the 16-stimulus continuum were prepared in a number of randomizations on magnetic tape. The 33 subjects who participated in the identification tests were told to choose the pair containing different members in each trial. The stimuli within each pair were separated by 250 msec, the pairs by one second, and the trials by three seconds. After every ten trials, there was an interval of seven seconds.

RESULTS

The identification data are displayed in Figure 1. The stimuli are arranged along the abscissa, and the percent identification along the ordinate with the functions showing, from left to right, responses as the low, mid or high tone. Although there is extensive overlap and the peaks do not exceed 92, 74 and 82 percent respectively, the results are clearly systematic. There were 1647 responses to each stimulus. To avoid clutter, I have not plotted the infrequent and unsystematic choices of the falling and rising tones as

---

[1]The remaining two possibilities are /khâa/ (falling) 'to kill' and /khǎa/ (rising) 'leg'.

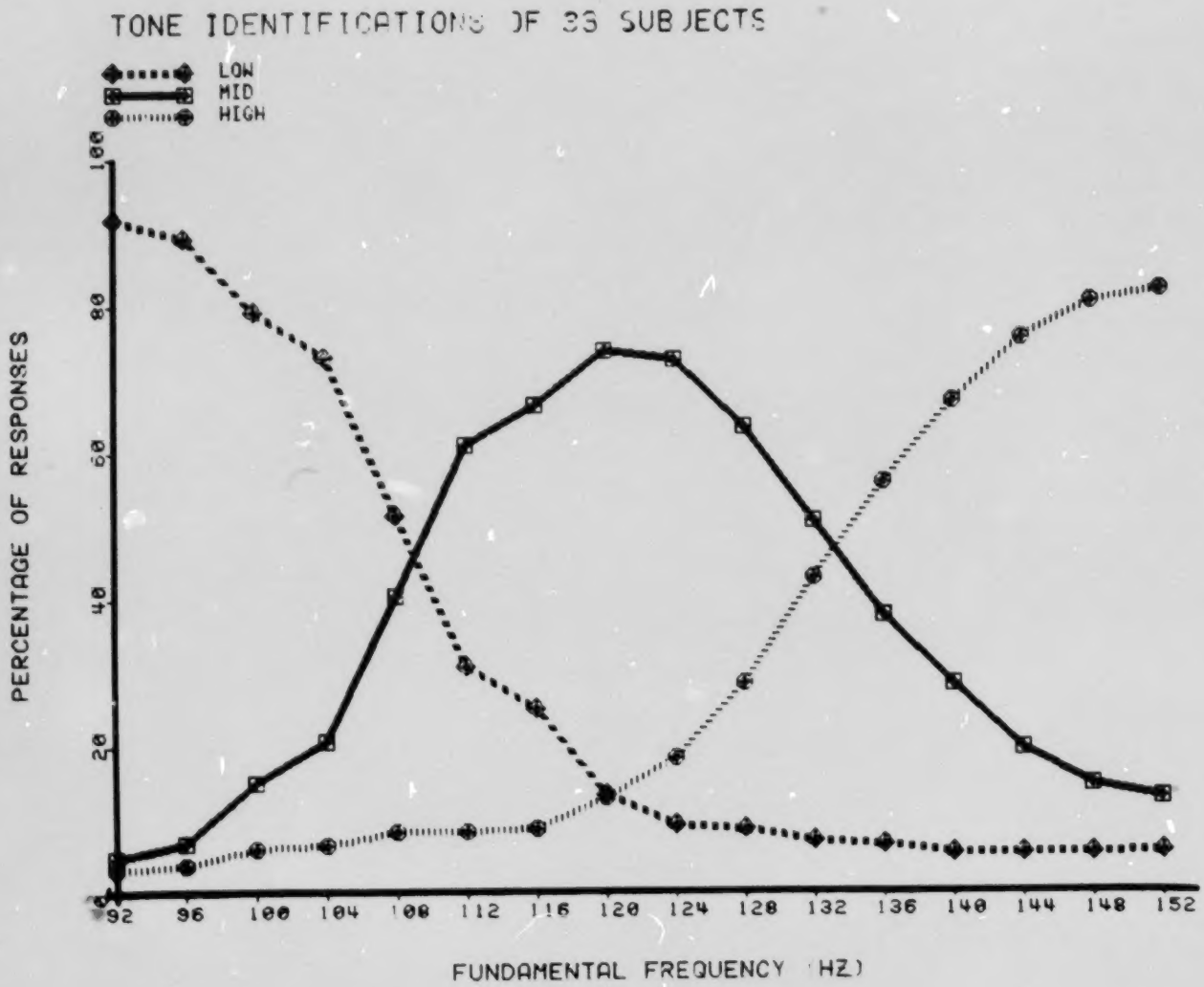

Figure 1: Labeling of 16 F_0 levels as the low, mid and high tones by Thai subjects.

response categories.

The 4IAX discrimination data for the 33 subjects are shown in Figure 2. Once again, the abscissa shows the array of stimuli. The ordinate gives the percent correct discrimination. The vertical lines at about 108 Hz and 133 Hz represent the crossover points of the identification curves in Figure 1 that mark the boundaries between the tonal categories as indicated. There were 1024 responses to each comparison. The lower function, representing the one-step discrimination, fluctuates between 80 and 90 percent with the lower part of that range favored at the high-frequency end. The two-step discrimination, represented by the solid line above, fluctuates between 92 and 96 percent. Customarily, predictions of acuity of discrimination are calculated from the identification functions. To do so here, however, would have been a fruitless exercise. Both levels of discrimination are so far above 50 percent, the chance level at the bottom of the graph, and so obviously unaffected by the category boundaries that there was really nothing to be demonstrated by such a calculation. Nevertheless, before coming to any final conclusion, I inspected the data more closely for signs of categoricity.

Within my expectation that some subjects would find the dimension more acceptable than other subjects did, I looked for a subset of people with especially good identification categories. I defined a "good" category as having an identification peak of at least 80 percent. The identification data for the 15 "good" subjects out of the original group of 33 are displayed in Figure 3 in which the peaks for the three tones are seen to achieve 98, 86 and 98 percent respectively. There were 771 responses to each stimulus. There is still considerable overlap, but it is less striking than before. The discrimination data for the same subjects are given in Figure 4, with 448 responses to each comparison. These curves differ only in minor and unsystematic ways from the curves for all 33 subjects in Figure 2. Although their data are not displayed separately, I did find five subjects who identified the tonal continuum in terms of just two categories, high and low, or mid and low. Their discrimination data did not differ significantly from any of the functions shown. Indeed, not one subject in all 33 showed boundary effects.

DISCUSSION

The results of my experiments are consistent with the view that the perception of Thai tonal categories is not categorical.[2] My results of 1961, based on a continuum of changing fundamental-frequency shapes yielding judgments of mid and high tone, are reaffirmed here by work with a continuum of flat variants yielding three tones as labeled by a much larger number of subjects, who also showed no effects of the phoneme categories in their discrimination performance. The discrepancy between Chan, Chuang and Wang and me (Abramson, 1961) remains to be explained. They cite Klatt (1973), who

---

[2]The title of a recent study by Siegel and Siegel (1977) implies that the perception of musical intervals, at least by musicians, is categorical, just as in certain phonetic dimensions. The article itself, however, makes it clear that there was no psychoacoustic testing of hearers' ability to discriminate variants of categories.

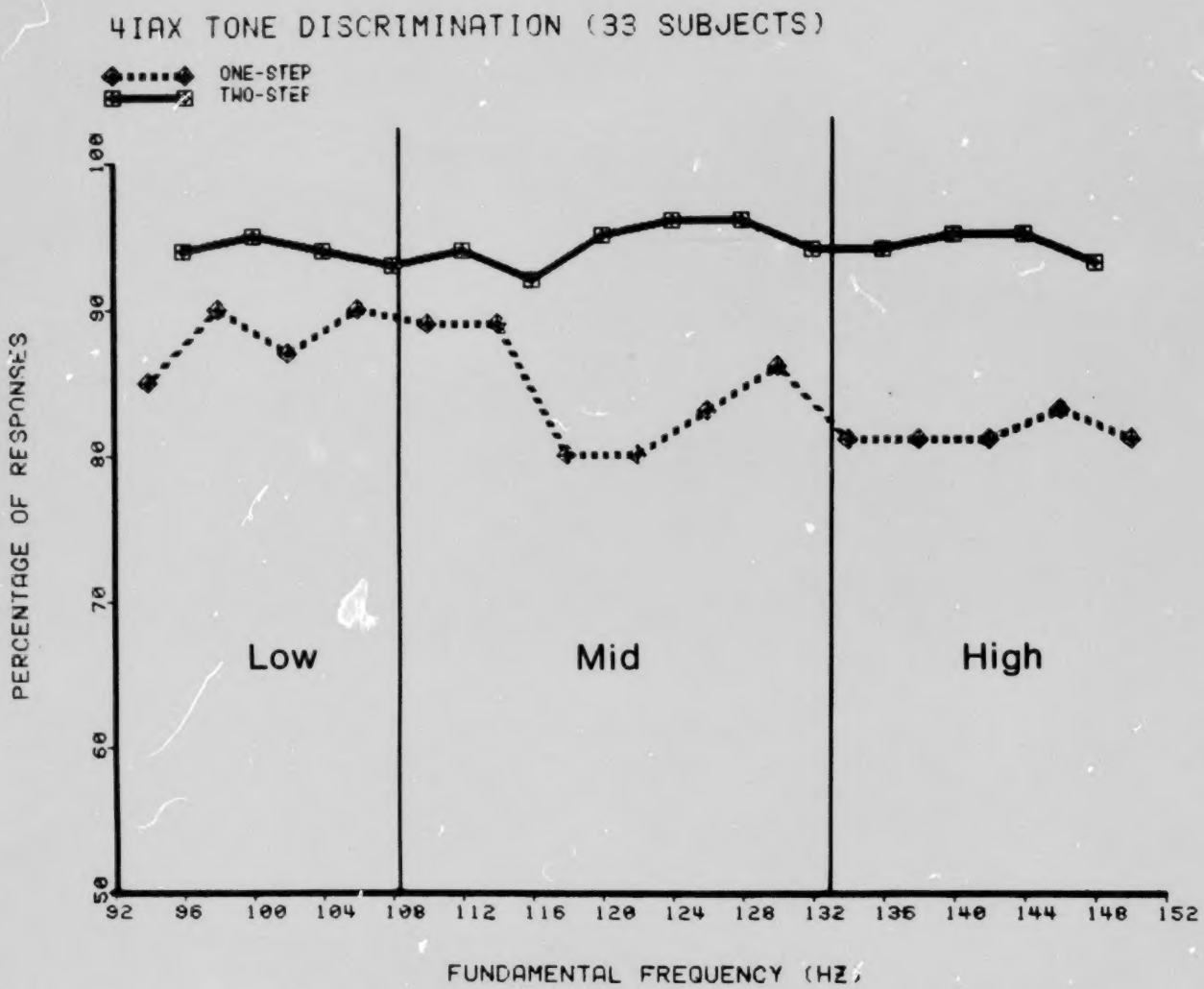

Figure 2: Discrimination of level-F_0 tonal variants by Thai subjects.

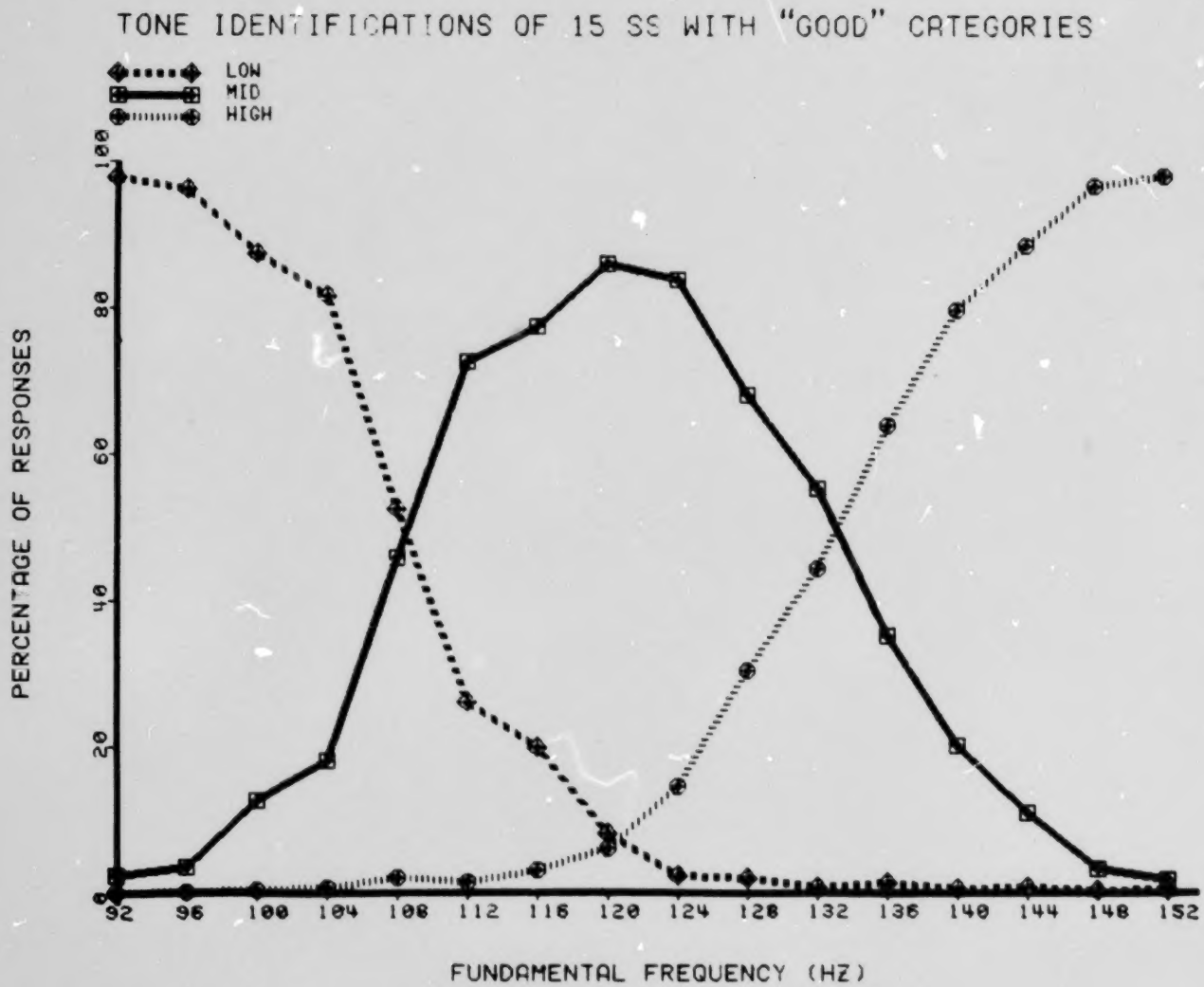

Figure 3: Labeling of F_0 levels by Thai subjects with "good" categories reaching at least 80 percent.

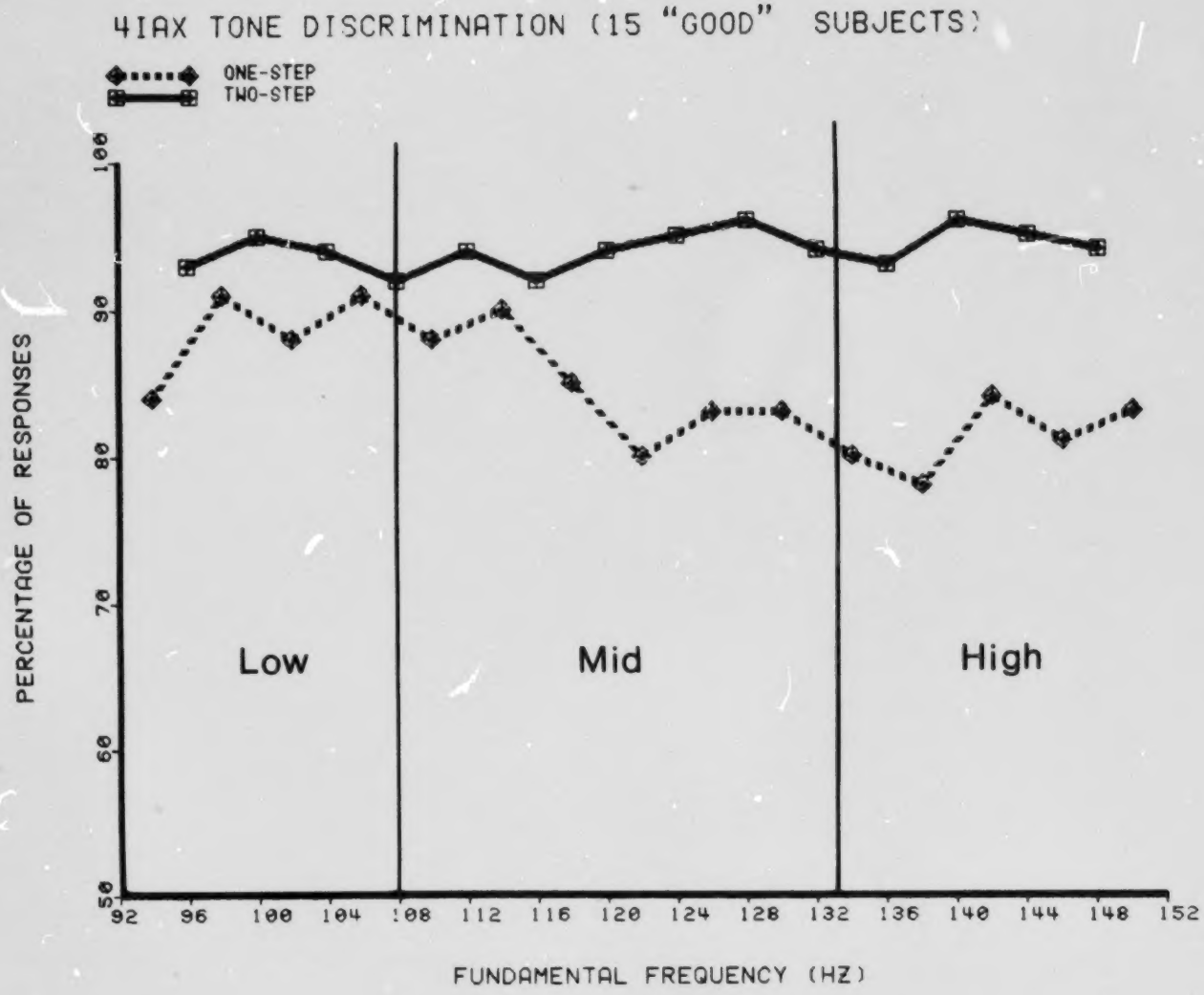

Figure 4: Discrimination of level-F_0 tonal variants by Thai subjects with "good" categories.

found that auditory sensitivity is greater to differences between unchanging fundamental-frequency contours than to differences between changing contours. Ten of the eleven variants in the Chan et al. study were rising contours,[3] while the eleventh, at the upper end, was flat. This might be a partial explanation of the difference. Differences between Mandarin Chinese and Thai may be relevant. Another difference is that Chan et al. synthesized the isolated vowel /i/ to yield two tonally differentiated words, while I used the somewhat more complex syllable [khaa] to yield three words. In any event, I have presented a continuum of rising and falling ramps elsewhere (Abramson, 1976) on the same syllable type that Thai subjects divided into the three tones of the present study. Perhaps a discrimination experiment with those variants will help in reconciling the two studies.

REFERENCES

Abramson, A. S. (1961) Identification and discrimination of phonemic tones. J. Acoust. Soc. Am. 33, 842(A).

Abramson, A. S. (1962) The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments. (Bloomington, Indiana: Indiana U. Res. Center in Anthropology, Folklore and Linguistics, Pub. 20).

Abramson, A. S. (1976) Static and dynamic acoustic cues in distinctive tones. J. Acoust. Soc. Am. 59, S542(A).

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967. (Prague: Academia), pp. 569-573.

Bastian, J. B. and A. S. Abramson. (1962) Identification and discrimination of phonemic vowel duration. J. Acoust. Soc. Am. 34, 743-744(A).

Chan, S. W., C. K. Chuan and W. S-Y. Wang. (1975) Cross-language study of categorical perception for lexical tone. J. Acoust. Soc. Am. 58, S119(A).

Erickson, D. M. (1976) A Physiological Analysis of the Tones of Thai. (Ph.D. Dissertation, The University of Connecticut).

Fry, D. B., A. S. Abramson, P. D. Eimas and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute (Univ. of Tokyo) 28, 67-73.

Klatt, D. H. (1973) Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. J. Acoust. Soc. Am. 53, 8-16.

Liberman, A. M., F. S. Cooper, D. S. Shankweiler and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., K. S. Harris, H. S. Hoffman and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

---

[3] In fact, their rising variants were a bit more complex in pattern. They remained flat at the starting F_0 for 100 msec and then rose to the final frequency value.

Pisoni, D. B. (1971) On the Nature of Categorical Perception of Speech Sounds. (Ph.D. Dissertation, The University of Michigan) [Suppl. to Haskins Laboratories Status Report on Speech Research].
Siegel, A. and W. Siegel. (1977) Categorical perception of tonal intervals: Musicians can't tell sharp from flat. Percep. Psychophys. 21, 399-407.

Effect of Speaking Rate on Vowel Formant Movements

Thomas Gay†

ABSTRACT

The purpose of this experiment was to study the effects of changes in speaking rate on both the attainment of acoustic vowel targets and the relative time and speed of movements toward these presumed targets. Four speakers produced a number of different consonant-vowel-consonant (CVC) and consonant-vowel-consonant-vowel-consonant (CVCVC) utterances at slow and fast speaking rates. Spectrographic measurements showed that the midpoint formant frequencies of the different vowels did not vary as a function of rate. However, for fast speech the onset frequencies of second formant transitions were closer to their target frequencies, while consonant-vowel (CV) transition rates remained essentially unchanged, indicating that movement toward the vowel simply began earlier for fast speech. Changes in both speaking rate and lexical stress had different effects. For stressed vowels, an increase in speaking rate was accompanied primarily by a decrease in duration. However, destressed vowels, even if they were of the same duration as quickly produced stressed vowels, were reduced in overall amplitude, fundamental frequency, and to some extent, vowel color. These results suggest that speaking rate and lexical stress are controlled by two different mechanisms.

INTRODUCTION

Within certain limits, speech perception does not appear to be constrained by rate of speech production; the information-bearing elements of segmental units are preserved across a wide range of speaking rates. Does the perceptual mechanism adapt to a different acoustic representation of these elements during fast speech, or are the articulatory gestures reorganized to produce a constant acoustic output? Some physiological evidence exists to

---

†Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research was carried out while the author was on leave at the Department of Speech Communication, Royal Institute of Technology, and Department of Linguistics, Stockholm University, Stockholm, Sweden. The comments and suggestions of Professors Gunnar Fant and Bjorn Lindblom of these institutions are gratefully acknowledged. This research was supported, in part, by grants from the National Science Foundation (BNS-7616954) and the National Institute of Neurological and Communicative Disorders and Stroke (NS-10424).

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

suggest the latter. In a series of experiments [Gay and Hirose (1973); Gay, Ushijima, Hirose and Cooper (1974); Gay and Ushijima (1975)] it was shown that the motor patterns underlying articulatory movements for fast speech were not only different from those during slow speech, but were reorganized in complex ways. In general, electromyographic activity associated with tongue body movements during vowel production decreased during fast speech, while activity associated with both labial and alveolar stop consonant production increased with an increase in speaking rate. At the movement level, while it would appear that vowel targets are not always reached during fast speech (Gay, et al., 1974), the tradeoffs between articulatory displacement and velocity seem to vary for individual speakers (Kuehn and Moll, 1976).

While it is apparent that changes in both motor programming and articulatory movements occur for changes in speaking rate, it is not known how these changes are reflected in the acoustic signal. Are acoustic targets the same for speech produced at slow and fast rates, or are these presumed targets systematically centralized, or otherwise shifted in frequency, as a function of rate? What are the temporal properties of consonant vowel (CV) transition movements for different speaking rates; do onset frequencies and rates of transition movements change for fast speech? The experiment reported in this paper was designed to study these questions by mapping the acoustic vowel space of several speakers across changes in speaking rate. A second purpose of the experiment was to study the acoustic effects of changes in speaking rate in relation to those for lexical stress to determine whether the two features can be accounted for by the same duration control mechanism.

METHOD

Subjects and Speech Material

Speakers were four adults, three males (WE, TG, LR) and one female (KH), all native speakers of American English. Two (TG, LR) spoke a New York dialect, one (KH) a New England dialect and one (WE) a General American (West Coast) dialect. While all four speakers were phonetically trained and experimentally sophisticated, none, except the author, knew the specific research goals.

Three different types of speech samples were constructed. The main set consisted of consonant vowel consonant (CVC) syllables containing the nine vowels, /i ɪ ɛ æ ɑ ɔ ʊ u ʌ/, in a /p_p/ environment. This frame was used for two reasons: one was that it paralleled that of an earlier physiological experiment (Gay et al., 1974), and the other is that it would probably produce minimal contextual effects. A second set consisted of a corresponding CVC subset with the point vowels, /i a u/ in a /b_p/ environment, and a third consisted of CVCVC sequences of the type, /ˈkipapˈ/, /kiˈpap/, /kapipˈ/, /kaˈpip/. The second set was used to provide a voicing contrast to parallel points of the main set, while the third was used to study the effects of changes in both speaking rate and lexical stress on the same syllable types. The sixteen speech samples were arranged, randomly, within each set, into a list. Each utterance was embedded in the carrier phrase, "It's a ___ again." Five such lists were constructed, one for each of five repetitions by each speaker.

Data Recording

The original protocol called for three different speaking rates to be used: one slow, or normal, and two fast. However, none of the speakers could comfortably or reliably produce speech at two different "fast" rates. An additional rate, slower than that of the normal one, was also considered, but it, too, was unnatural and resulted in obviously contrived renditions. Apparently, these four speakers, at least, have two natural rate modes, one normal (or slow) and one fast, with rates outside those two being difficult to control or maintain. Thus, only two rates were studied, both determined by each speaker's own judgment of natural slow and fast rates. Each speaker read each of the five utterance lists through, first at the slow rate and then at the fast rate. For the CVC utterances, subjects were instructed to place sentence stress on the test word. For the stress contrasts, the speakers were instructed to destress the appropriate syllable while still maintaining its phonetic identity; in other words, not to the point where the vowel would be reduced in color to a schwa. All subjects received detailed instructions about the tasks before the recording session began, and had ample practice time with the utterances. All recordings were made in a sound-treated room.

Data Analysis

A total of 640 utterances were analyzed (16 samples x 5 repetitions x 2 rates x 4 speakers). Spectrograms were made for each utterance on a Voiceprint Laboratories Sound Spectrograph, using the extended frequency scale, wide band filter (300 Hz), and highshaping setting. In addition, fundamental frequency and overall amplitude measurements were made for the stress contrasts using a specially modified pen writing oscillograph (Mingograf, 34T). Duration and formant frequency measurements were made from the spectrograms at points indicated in Figure 1. Duration measurements were made for the stop gap closure of the initial consonant, CV transition, a combined measurement of the vowel nucleus and, if present, the final VC transition, and closure for the final consonant. The VC transition component was included in the vowel nucleus measurement because it was difficult to segment out. Formant frequency measurements (F_1, F_2, F_3) were made at the time of release of the initial consonant, the vowel midpoint, and at the time of closure for the final consonant. If the CV transition was not visible at the time of consonant release, its position was straightline extrapolated. The vowel midpoint was defined in one of three ways: 1) the point where the F_2 transition reached a steady state, 2) if a steady state was not reached, the point where the F_2 transition reached maximum displacement before changing direction, or 3) if the transition was unidirectional, at a point midway between the onset of voicing and closure for the final consonant. Most of the vowels met criteria 1 or 2; however, for two speakers, both /u/ and /ɔ/, consistently, and /I/ and /ɛ/, occasionally, were characterized by unidirectional glides from initial consonant to final consonant. The method of tracking and locating the measuring points was the usual one: a pencil line was drawn through the center of both the transition and steady state (where present) portions of the formant; measurements were made where the line intersected the point of consonant release (for the transition onset) and at either the center of the steady state portion or where the line intersected that of the CV transition.

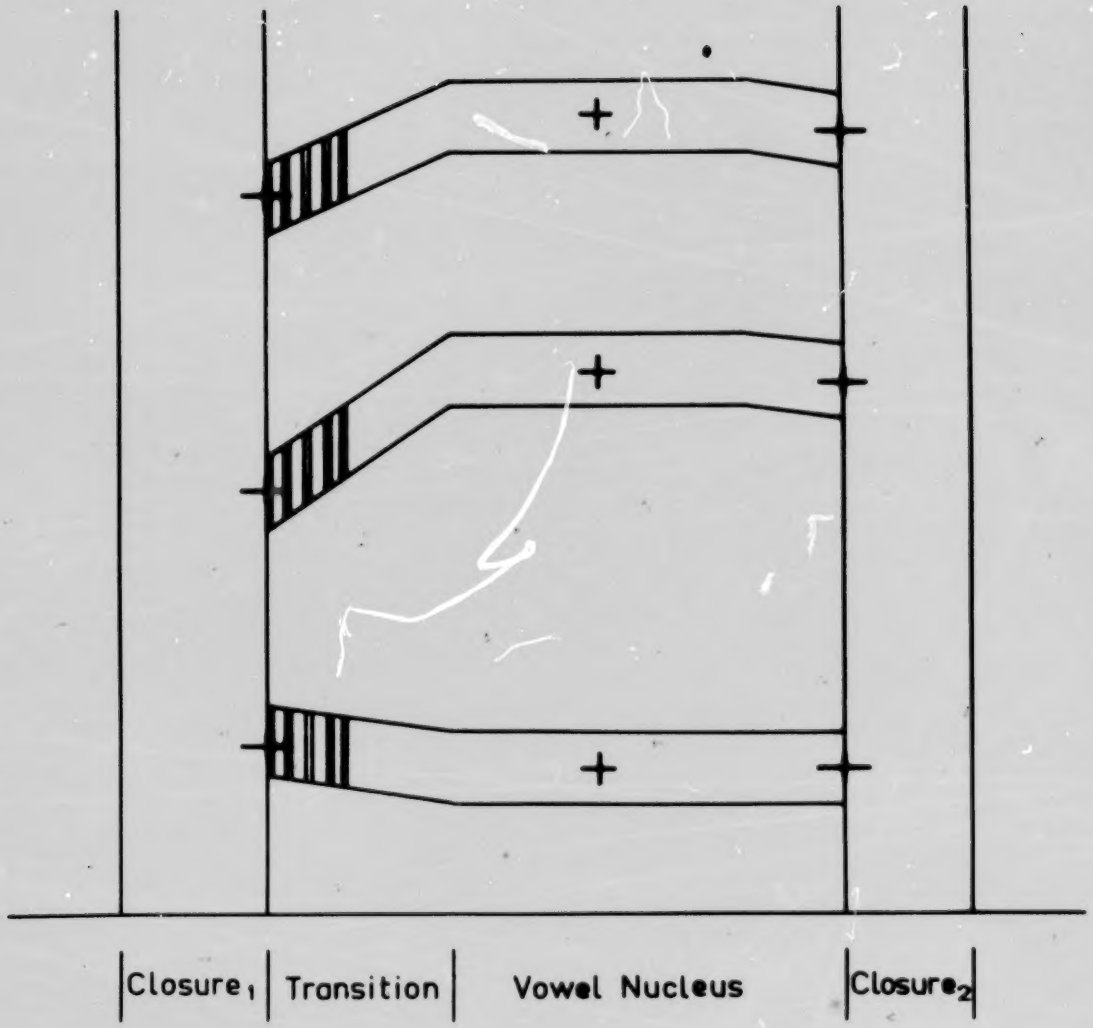

Figure 1: Schematic illustration indicating points where duration and formant frequency measurements were made.

First and second formants were visible for all subjects, but F_3 did not always appear clearly for Speakers TG and LR, and for most samples of the destressed syllables. As would be expected, CV movements for F_1 were small, while those for F_2 provided the most reliable and useful transition movement information. Repeated measurements of selected samples revealed fairly consistent error ranges. Duration measurements were accurate to $\pm$ 10 msec, while formant frequency measurements were accurate to $\pm$ 25 Hz. The latter range is consistent with that of other reports (Lindblom, 1961; Öhman, 1965a). For the most part, the token-to-token variability for each subject was small. Durations (as measured from closure of the initial consonant to release of the final consonant) usually varied by no more than 25 msec within each rate compared to an average of 45 msec for across-rate differences. Similarly, token-to-token formant frequency variations usually fell within the range of error measurement ($\pm$ 25 Hz). This stability might be due to the highly structured nature of the utterances and the stress pattern of the sentence. There were, of course, some exceptions to this general stability; these exceptions will be discussed in the following section.

RESULTS

The effects of differences in speaking rate on the durations of each of the segments of the main set of CVC syllables are summarized in Table 1. The purpose of this table is to show how the overall decrease in syllable duration during faster speech is absorbed by each of the constituent segments. The table shows the durations of initial /p/ closure, CV transition, vowel nucleus, and final /p/ closure, pooled over both utterance repetitions and speakers. Each value represents the mean of twenty (5 repetitions x 4 speakers) tokens. While the actual durations of the different syllables varied for the four speakers, relative differences, both among vowels and between rates, are represented in the means.

The table shows, not unexpectedly, that the reduction in duration during fast speech is reflected primarily in the duration of the vowel, although perhaps not to as great a degree as might be intuitively expected. Differences in the durations of the vowel nuclei for slow and fast speech ranged from 20 to 35 msec, depending on the particular vowel. While the overall duration of the vowel varied from 105 msec for /i/ to 165 msec for /ɔ/ at the slow rate, the percentage change did not vary as a function of duration. For example, the phonetically long vowels /æ/ and /ɔ/ were not reduced to any greater or lesser percent during fast speech than the phonetically short vowels, /I/ or /ʊ/. The slow rate vowel durations obtained in this experiment are considerably shorter than those reported by Peterson and Lehiste [(1960), 245 msec] for CVC words in a similar carrier, but slightly longer than those reported by Klatt [(1975), 110 msec] for vowels spoken in connected discourse. These differences are probably related to differences in phonetic context as well as utterance position in the sentence carrier.

Although the vowel nucleus absorbed most of the decrease in duration during fast speech, the consonant segments were also consistently shorter as well. However, differences in duration for initial and final consonant closure were considerably less between the two rates; also, greater variability and even some overlap occurred in certain instances at the

TABLE 1: Mean durations of consonant closure (1), vowel, and consonant closure (2), pooled over all repetitions and speakers. For each vowel, values for the slow rate appear on the first row, those for the fast rate on the second row.

	Consonant Closure (1)		Vowel		Consonant Closure (2)	
	duration	ratio	duration	ratio	duration	ratio
i	100	.95	120	.75	80	.88
	95		90		70	
I	105	.90	105	.81	80	.94
	95		85		75	
ɛ	105	.90	130	.81	90	.89
	95		105		80	
æ	105	.90	155	.81	80	.88
	95		125		70	
a	100	.90	145	.79	80	.94
	90		115		75	
ɔ	105	.95	165	.79	80	.88
	100		130		70	
ʊ	105	.95	110	.82	85	.88
	100		90		75	
u	100	.95	120	.75	80	.81
	95		90		65	
ʌ	100	.90	115	.74	80	.94
	90		85		75	
mean	105		130		80	
	95		100		75	

individual token level. Prestressed initial /p/ is consistently, and usually substantially, longer than poststressed final /p/, for both rates, as expected. Interestingly, although the vowel portion is most affected during fast speech, the contributions of the initial and final consonants account for at least one-third of the total reduction in syllable duration. It was also found that the transition durations within each rate were relatively stable across the different vowels. However, transition time was reduced somewhat during fast speech, to about the same degree as that for consonant closure, some 5-10 msec. Transition times ranged from 40-50 msec for the slow rate and 35-45 msec for the fast rate. The stable transition times across vowels are consistent with the articulatory data of both Kent and Moll (1969) and Kuehn and Moll (1976), but shorter and less variable than those reported by Lehiste and Peterson (1961) and Öhman (1965b). These differences might be due to differences in overall duration, phonetic context, and carrier structure.

The major question of interest in this paper is whether the acoustic targets of vowels (as measured at the midpoint) vary, either systematically or unsystematically, as a function of speaking rate. Spectrographic measurements of first, second, and third formant frequencies show that they do not. Figure 2 shows the F_1-F_2 vowel space for all nine vowels produced by each speaker. Each data point represents the mean of the five repetitions by that speaker at each rate. The overall picture is one of little variability. For the most part, the means for each rate (both F_1 and F_2) are quite close, usually falling well within the range of error measurement. However, some instances of variability between the slow and fast rates do occur. For Speaker WE, F_1 for /æ/ is approximately 50 Hz lower for the fast rate. This difference, statistically significant at the .02 level of confidence (t test for independent means), might be explained by jaw undershoot during fast speech; however, if this is true, it is curious why the other open vowels are not similarly affected. Speaker LR also showed some front vowel variability, but in this case, only F_2 for /ɛ/ is significantly different (.01 level of confidence). The greater variability between rates appears in the data for the single female speaker (KH). While the range of variability is still small, it is nonetheless greater than that for the other speakers. In addition, variability within each rate is greater for this speaker than for any of the others. The wider range of variability for this speaker might be a simple consequence of the increased probability of error encountered when measuring female formant frequencies, a speculation that is supported by the fact that only the means of F_1 for /æ/ were significantly different (.05) as a function of rate.

While vowel durations for three of the four speakers were distributed essentially bimodally between the two speaking rate conditions, one speaker (TG) often showed considerable temporal variability, with vowel durations distributed along a continuum. One such instance was for the vowel /i/. Figure 3 shows the first, second, and third formant frequencies (measured at the midpoint) for /i/ plotted as a function of duration (transition + nucleus). The twenty tokens represent all of those produced at both rates for both the /p/ and /b/ syllables for this speaker. It is apparent from this figure, that the effect of speaking rate on the attainment of acoustic vowel targets is negligible, even over a wide (55 msec) range of durations. The ranges of formant frequency variations are 50 Hz for F_1, 75 Hz for F_2, and 75 Hz for F_3. It should be noted that these midpoint frequencies are attained

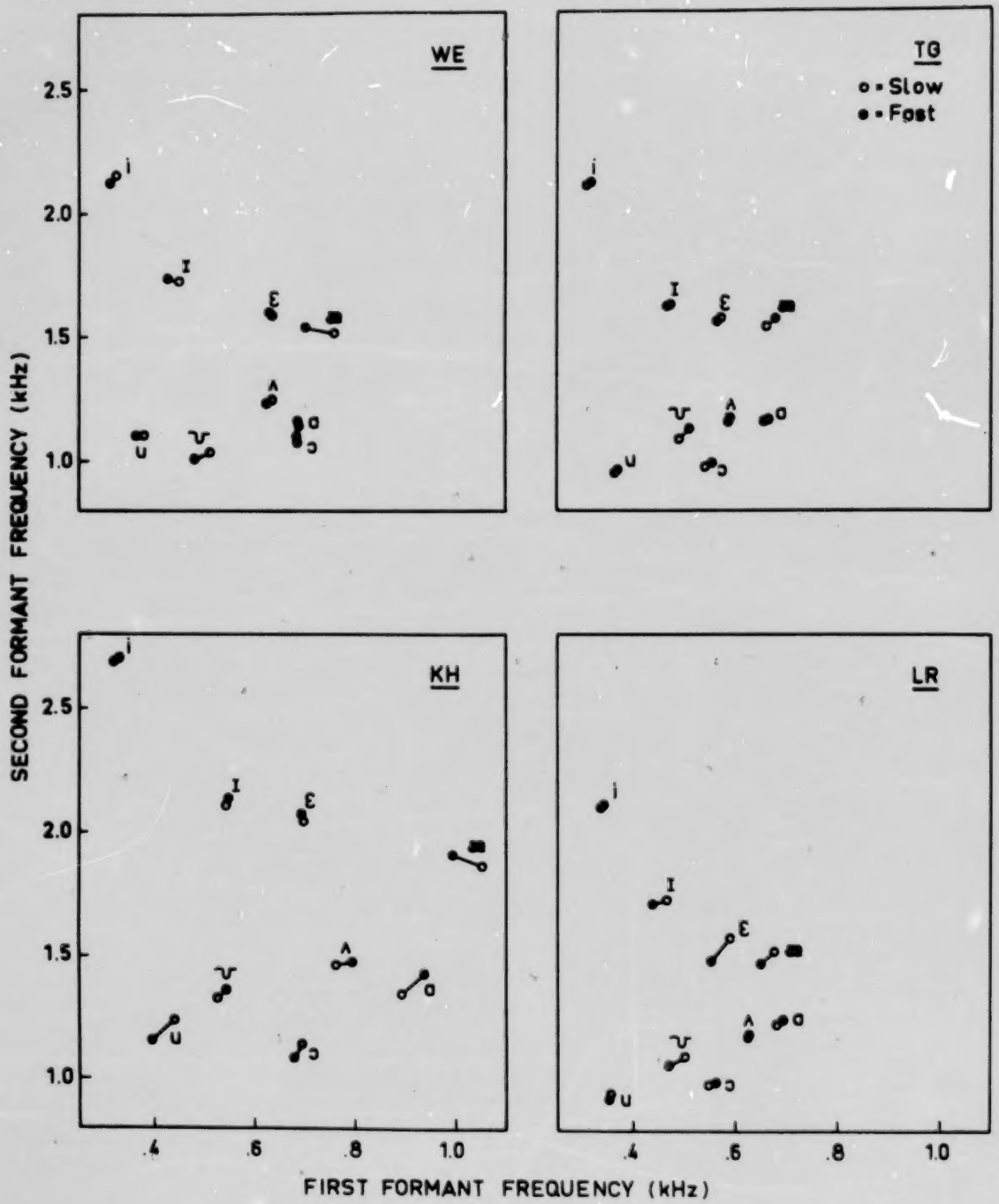

Figure 2: $F_1 - F_2$ vowel space for midpoint measurements for both speaking rates.

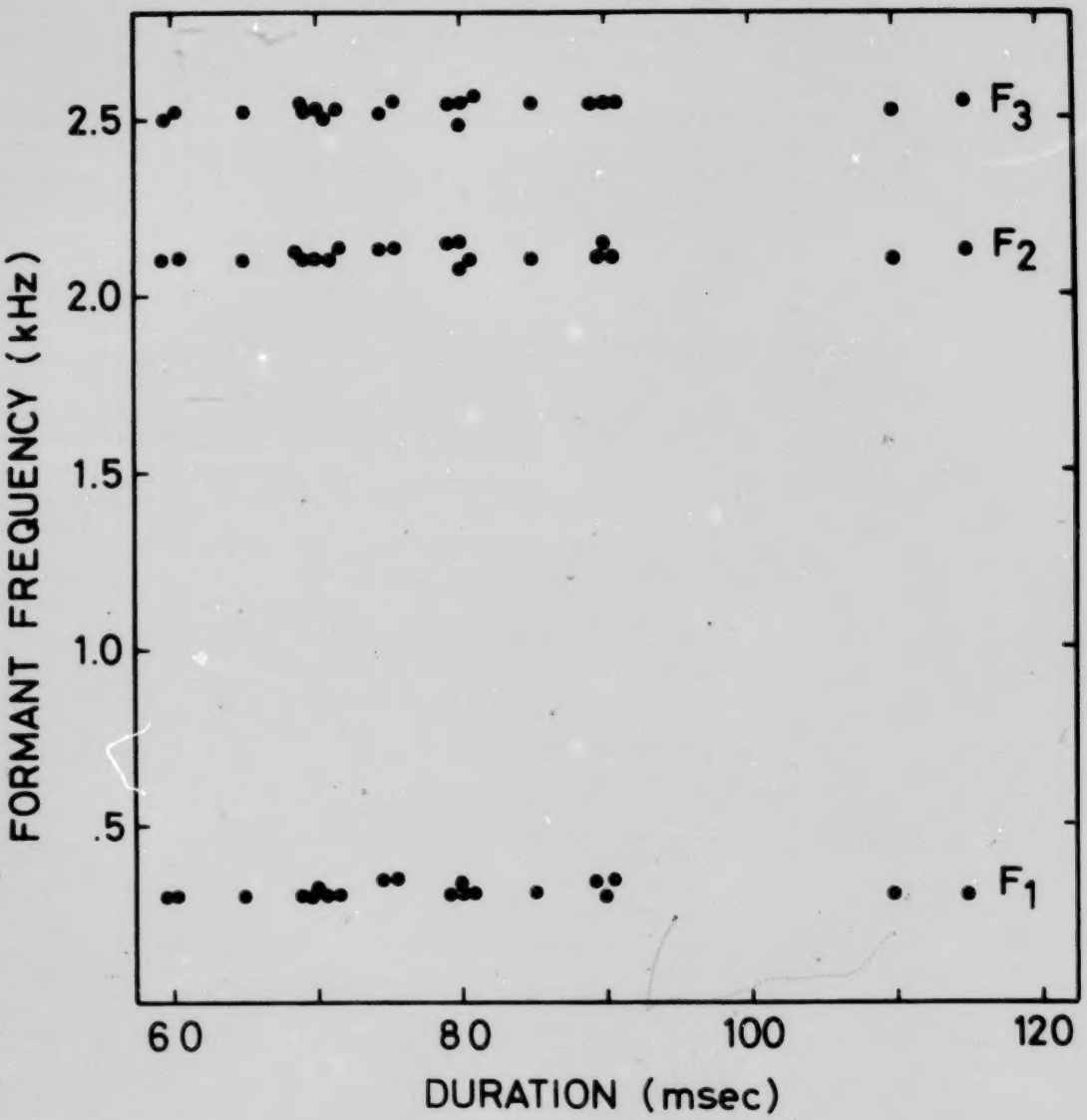

Figure 3: Midpoint frequencies for F_1, F_2, F_3, as a function of duration, for the vowel /i/ (Speaker TG).

through CV transitions that originate, at consonant release, some 300-400 Hz lower in frequency. It is also apparent that the observed variability is not correlated with duration along any part of the continuum.

Assuming that the same acoustic target is reached for a vowel spoken at both slow and fast rates, the question becomes how the underlying articulatory gesture is modified to achieve that constant end. Either of two types of adjustments seems likely: first, articulator movement toward the vowel target can begin earlier in time, that is, closer to the time of initial consonant closure, or second, the movement can be produced at a faster rate of speed. Evidence for the first can be seen spectrographically by a change in the position of the onset frequency of the CV transition as a function of rate; specifically, the onset frequency of the formant transition would be closer to that of the target frequency for faster speech. An estimation of the second could be made by simply calculating the overall velocity of the formant transition itself. These measurements appear for the vowel /i/ for all speakers in Table 2. This table shows the mean durations for /p/ closure, the CV transition and the vowel nucleus, and the F_2 onset and midpoint frequencies and F_2 rate of change, for both speaking rates. The measures are pooled over utterance repetitions. The F_2 rate was calculated by dividing the difference between the F_2 midpoint and F_2 onset frequencies by the transition duration.

TABLE 2: Duration and formant frequency measurements for the vowel /i/, pooled over utterance repetitions.

	Duration			Frequency		
Speaker (S/F)	Closure	Transition	Vowel	F_2 onset	F_2 midpoint	F_2 rate
WE	95	45	115	1925	2150	4.5
	80	40	95	1965*	2125	4.3
TG	95	50	95	1765	2105	7.7
	90	50	75	1845***	2115	6.9
KH	105	55	140	2230	2700	8.8
	110	45	105	2300	2685	8.9
LR	110	50	130	1735	2100	7.7
	105	40	95	1780**	2115	8.4

*Significant at .10
**Significant at .05
***Significant at .01

For all subjects, the onset frequency of the second formant transition is higher for the fast rate condition, while the F_2 midpoint frequencies and F_2 rates of change remain essentially unaffected across the two rates. The differences in mean onset frequencies are statistically significant (at various levels of confidence) for all but the female speaker. The mean differences range from 40 Hz for subject WE to 80 Hz for subject TG. The corresponding differences in midpoint frequencies, however, are only of the order of 15-25 Hz for all speakers. It should also be noted that these shifts occur in syllables whose closure durations are shorter, a condition which, theoretically at least, would tend to minimize the observed effects.

A difference in second formant transition onset frequencies between slow and fast speech is also evident where /b/ is the initial consonant. Measurements for /b/ appear in Table 3, which contains essentially the same data as in Table 2 but for the /pip-bip/, /pap-bap/ and /pup-bup/ contrasts, pooled over the three male speakers. For these three vowels, at least, the rate effects that exist for /p/, exist for /b/ as well, except perhaps to a slightly lesser extent. Again, for all three vowels preceded by /b/, the onset frequency of the second formant transition is closer to the midpoint frequency, while the midpoint frequencies, themselves, and the F_2 rates remain largely unaffected. The difference in onset frequencies is slightly less for /b/ than /p/. For /bip/, the shift across rates is 25 Hz, while for /pip/, the shift averages 60 Hz. For /a/ and /u/, however, the shifts are more comparable.

TABLE 3: Duration and formant frequency measurements for voiced-voiceless contrasts. Values are pooled over all repetitions for the three male speakers.

Utterance (S/F)	Closure Duration	Transition Duration	F_2 onset	F_2 midpoint	F_2 rate
pip	100	50	1810	2120	6.6
	90	45	1885	2120	6.5
bip	105	45	1960	2120	3.6
	95	40	1985	2105	3.2
pap	95	50	1340	1170	2.1
	85	45	1325	1180	2.0
bap	105	-	1130	1150	-
	95	-	1160	1170	-
pup	95	40	1250	1000	3.6
	85	40	1200	990	3.9
bup	100	45	1065	990	0.8
	90	35	1030	990	1.2

More obvious than the rate differences are the differences in second formant transition onset frequencies and rates of change between /p/ and /b/, within each speaking rate condition. For all three vowels, it is apparent that the onset frequencies are considerably closer to the midpoint frequencies, and the F_2 rate is correspondingly slower, for /b/ as opposed to /p/. All frequency differences between /p/ and /b/ were statistically significant for all four speakers at either the .05 and .01 level of confidence. The absence of a CV transition for /bap/ indicates that the movement towards /a/ from /b/ was probably completed before release of the consonant occurred, much earlier than the corresponding transition from /p/. The greater range of F_2 onsets for /b/ as opposed to /p/ is consistent with Fant's (1969) calculations for similar CV syllables; even the extent of the differences is similar. One explanation for this frequency shift is that because of the greater tenseness associated with /p/, the tongue is not as free to move toward the vowel as it is for the more lax /b/; in other words, the tongue is more free to coarticulate with the following vowel during /b/ than it is during /p/. A second possibility is that the voiceless consonant /p/ occurs earlier in time during the vowel-to-vowel movement than /b/, producing transitions that are temporally offset to the left. It is also conceivable that part of the frequency difference between /pap/ and /bap/ is due to the presence of an additional subglottal formant associated with /p/ release [Fant, Ishizaka, Lindqvist and Sundberg (1972)]. Interestingly, then, while changes in phonetic context affect the pattern of movements toward the vowel target, the frequencies of the targets, themselves, do not appear to be affected, nor is the basic rate effect on the movements different. The rate effects seem to be superimposed on the context-dependent articulatory movements, and are not affected by, or assimilated into, these movements.

The final set of measurements is related to the question of how changes in both speaking rate and lexical stress affect the acoustic properties of vowels. Are the effects of these two features, both of which affect vowel duration, additive or independent? The duration and frequency measurements for both the /i/ and /a/ stress contrasts appear in Tables 4 and 5. These tables show the measurements of vowel duration, relative overall amplitude, fundamental frequency, and first and second formant frequencies for the second syllable of the utterance for each speaker. All frequency and amplitude measurements were made at the vowel midpoint. The amplitude measurements are in dB relative to the least intense utterance (= 0) for each speaker. All values are pooled over the five utterance repetitions.

For both /i/ and /a/, a number of differences appear between the stressed and unstressed pairs for each speaking rate, but virtually no differences emerge between rates for the same stress condition. The slow and fast pairs within each stress condition are characterized by essentially the same overall amplitude, fundamental frequency, and first and second formant frequencies. However, the corresponding unstressed syllables at each rate are consistently lower in overall amplitude and fundamental frequency, and somewhat reduced in vowel color. Fundamental frequency differences were statistically significant for all speakers (.01) for the stressed-unstressed pairs, while reduction of F_1 and F_2 was significant (.05) only for speaker KH.

The overall stress findings are, of course, consistent with those of a number of earlier studies [Fry (1955); Lieberman (1960); Lindblom (1963);

TABLE 4: Duration, relative amplitude, and frequency measurements for the vowel /i/ as a function of both stress and speaking rate.

	Stressed					Unstressed				
Speaker (S/F)	Duration	Rel. Amp.	F_0	F_1	F_2	Duration	Rel. Amp.	F_0	F_1	F_2
WE	140	6	140	300	2150	115	0	110	300	2125
	90	6	150	300	2120	95	0	110	325	2100
TG	90	6	125	315	2155	80	5	95	335	2085
	70	8	135	325	2125	70	0	95	330	2050
KH	125	7	225	340	2710	100	3	150	425	2530
	105	4	220	330	2670	75	0	160	430	2505
LR	115	10	140	315	2085	95	0	110	320	2090
	95	7	135	335	2120	80	0	110	330	2010

TABLE 5: Duration, relative amplitude, and frequency measurements for the vowel /a/ as a function of both stress and speaking rate.

	Stressed					Unstressed				
Speaker (S/F)	Duration	Rel. Amp.	F_0	F_1	F_2	Duration	Rel. Amp.	F_0	F_1	F_2
WE	150	4	140	665	1125	130	2	110	650	1150
	105	4	145	675	1150	85	0	115	640	1125
TG	120	3	110	675	1155	115	1	90	660	1190
	100	6	125	675	1165	100	0	95	625	1175
KH	145	5	230	910	1400	140	0	155	850	1450
	115	5	220	880	1380	95	1	165	800	1500
LR	155	4	140	665	1230	120	1	110	650	1240
	125	6	130	660	1250	100	0	110	600	1250

Brown and McGlone (1974)]. Of particular interest in the present data, however, is that these differences are apparently not related primarily to differences in duration. For example, while the "unstressed-slow" syllables are, in at least half the cases, roughly comparable in duration to the "fast-stressed" syllables, they are nonetheless considerably reduced in fundamental frequency, and somewhat reduced in overall amplitude and vowel color with respect to their "fast-stressed" counterparts. These data seem to indicate that while speaking rate and lexical stress both affect the duration of vowel segments, they have different effects on several acoustic parameters, and are probably independently controlled by different physiological mechanisms.

DISCUSSION

The results of this experiment show that differences in vowel duration due to changes in speaking rate do not seem to have a substantial effect on the attainment of acoustic vowel targets. The formant frequencies of these presumed targets remained essentially unchanged across changes in speaking rate. It was also shown that the probable mechanism by which these targets were achieved was an earlier onset of the transition movement from consonant to vowel. The speed of movement from the consonant to the vowel, however, did not seem to change. While these patterns were consistent across all five speakers, they were observed for only a small number of phonetic samples produced by phonetically trained speakers in a precise manner. Further, the present results might also be affected by several additional factors that were not studied in the present experiment. One such complicating factor might be differences in phonetic context. For example, the effect of an increase in speaking rate on transition movements might be different depending on whether the movement is from an alveolar or labial consonant. Likewise, because vowel duration is conditioned by factors other than speaking rate, differences in speech material, phonetic context, and word position in a sentence, for example, Klatt (1976), shorter segment durations associated with one of these factors might produce a different pattern of CV transition movement.

Differences in overall duration might account for the differences between the present results and the articulatory data of Kuehn and Moll (1976). Kuehn and Moll showed that different speakers can use different strategies to control speaking rate, with one such observed strategy being an adjustment of articulatory velocity. This obvious inconsistency between the two sets of data might be related to differences in corresponding across-rate durations. In the present experiment, transition durations for fast speech were approximately 90 percent of those for slow speech, while the corresponding fast speech durations measured by Kuehn and Moll were on the order of 50 percent of those for slow speech. Thus, it might be suggested that if changes in articulatory velocity (and corresponding transition rate of change) appear, they might do so primarily at very fast rates of speech.

The present acoustic data are also inconsistent with earlier EMG data [Gay, et al (1974); Gay and Ushijima (1975)] that showed a change in the level of muscle activity for vowels in response to a change in speaking rate. The EMG data showed that the activity levels of the genioglossus muscle for the vowel /i/ decreased with an increase in speaking rate. The genioglossus is a prime mover of the tongue and is active during, and probably responsible for, the bunching and protruding movement of the tongue for /i/. The decrease in

activity implies either, or a combination of both, a decrease in articulatory displacement or a decrease in the speed of articulatory movement. However, the present acoustic data show that neither of these parameters (as reflected in the acoustic signal) are substantially affected. The different interpretations that arise from the physiological and acoustic data might be explained in a number of ways, none of which seem entirely satisfactory. First, the reduction in EMG activity might not reflect a corresponding difference in articulatory displacement, that is, undershoot at the muscle contraction level might not produce undershoot at the articulator movement or acoustic level. Second, a totally different motor strategy using different muscles in different ways might come into play during fast speech. Third, the peak of the integrated EMG envelope might not provide an accurate indication of the maximum strength of contraction of an active muscle when the duration of the muscle contraction is changed. Changes in the peak of the EMG envelope are usually interpreted as reflecting (without being able to separate) changes in either the displacement of the articulator that the muscle is acting directly upon, or changes in the speed of movement of that articulator. However, it is also possible that the summated potentials of the integrated signal might peak differently if the duration of the contraction changes. Thus, with all other parameters held constant, a reduction in the peak of the integrated EMG envelope might also reflect simply a reduction in the contraction time of that muscle.

The inconsistencies between the physiological and acoustic data aside, it would appear from the data of this experiment that the coordination of articulatory movements is adjusted in some way in order to preserve the information bearing elements of segmental units across changes in speaking rate. The reduction in duration of all segments (ref. Table 1) coupled with the relative constancy of acoustic (vowel) targets, suggest that this adjustment involves primarily a horizontal compression along the time dimension. This type of compression, the existence of which was suggested some thirty years ago [Joos (1948)], is a nonlinear one, and one that causes both a decrease of duration within segments and an increase in coarticulation between segments. It also appears that temporal restructuring for changes in rate is superimposed on the basic serial ordering process.

The control of, and effects of changes in, speaking rate and lexical stress seem to be different in a number of ways. The data of this experiment show that for stressed vowels, only duration is reduced to any substantial degree. However, destressed vowels, even if they are of the same duration as quickly produced stressed vowels, are reduced in overall amplitude, fundamental frequency, and to some extent, vowel color.

The finding that a destressed vowel was not substantially reduced in color toward the neutral schwa does not completely coincide with either Lindblom's (1963) acoustic data or Harris' (1975) EMG data. These differences are probably due to the fact that in the present experiment, speakers were explicitly instructed to maintain the phonetic identity of the vowel during destressing. It might be suggested that if extended stress contrasts were studied in the present experiment, a greater degree of vowel reduction might have been observed. While differences in the degree of vowel reduction between Lindblom's (1963) findings and the present data can be explained, the question of the relationship between reduction and duration is more difficult

to resolve. Because vowel reduction appeared for changes in both stress and speaking rate in Lindblom's data, he concluded that reduction (and undershoot) was caused solely by changes in duration, and not the suprasegmental features of stress and rate, per se. However, the present data lead to the opposite conclusion. Because the tendency for formant frequencies to be reduced toward the neutral schwa occurs only for an unstressed vowel, even if it is of the same duration as its stressed counterpart, the present data suggest that the degree of reduction is linked to stress, regardless of the relative or absolute duration of the segment. The suggestion that stress, and not duration, determines target attainment has also been put forth by both Harris (1975) and Nord (1975).

In this experiment, it was also shown that destressing affected the fundamental frequency and overall amplitude of the vowel, indeed, even more so than the formant structure. These effects, which are consistent with those described by Fry (1955) and Lieberman (1960) among others, are compatible with an "extra effort" model of stress, such as the one proposed by Öhman (1967). A reduction of overall articulatory effort can result in corresponding reductions in the four parameters measured in this experiment: fundamental frequency, overall amplitude, duration, and vowel color. Thus, the findings of this experiment suggest that two separate and independent physiological mechanisms control changes in speaking rate and lexical stress, one that horizontally compresses the string and the other that modulates overall articulatory effort. A change in duration is a deliberate strategy of the first, while only a consequence of the second.

REFERENCES

Brown, W. S. and R. McGlone. (1974) Aerodynamic and acoustic study of stress in sentence productions. J. Acoust. Soc. Am. 56, 971-974.

Fant, G. (1969) Stops in CV syllables. Speech Transmission Laboratory, QPSR 4/1969, R.I.T., 1-25.

Fant, G., K. Ishizaka, J. Lindqvist and J. Sundberg. (1972) Subglottal formants. Speech Transmission Laboratory, QPSR 1/1972, R.I.T., 1-13.

Fry, D. (1955) Duration and intensity as physical correlates of linguistic stress. J. Acoust. Soc. Am. 27, 765-768.

Gay, T. and H. Hirose. (1973) Effect of speaking rate on labial consonant production: a combined electromyographic high-speed motion picture study. Phonetica 27, 203-213.

Gay, T. and T. Ushijima. (1975) Effect of speaking rate on stop consonant-vowel articulation. Proc. Speech Comm. Seminar, 1974, ed. by G. Fant. (Stockholm: Almqvist and Wiksell), 205-209.

Gay, T., T. Ushijima, H. Hirose and F. S. Cooper. (1974) Effect of speaking rate on labial consonant-vowel articulation. J. Phonetics 2, 47-63.

Harris, K. S. (1975) Mechanisms of duration change. Proc. Speech Communication Seminar-74, ed. by G. Fant. (Stockholm: Almqvist and Wiksell), 299-305.

Joos, M. (1948) Acoustic phonetics. Lang., (S), 136.

Kent, R. and K. Moll. (1969) Vocal-tract characteristics of the stop cognates. J. Acoust. Soc. Am. 46, 1549-1555.

Klatt, D. H. (1975) Vowel lengthening is syntactically determined in a connected discourse. J. Phonetics 3, 129-140.

Klatt, D. H. (1976) Segmental duration in English. J. Acoust. Soc. Am. 59,

1208-1221.

Kuehn, D. and K. Moll. (1976) A cineradiographic study of VC and CV articulatory velocities. J. Phonetics 4, 303-320.

Lehiste, I. and G. E. Peterson. (1961) Transitions, glides, and diphthongs. J. Acoust. Soc. Am. 33, 268-277.

Lieberman, P. (1960) Some acoustic correlates of word stress in American English. J. Acoust. Soc. Am. 32, 451-454.

Lindblom, B. (1961) Accuracy and limitations of Sonagraph measurements. Proc. Fourth Int. Cong. Phonetic Sciences. (s'Gravenhage: Mouton), 208-213.

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Nord, L. (1975) Vowel reduction: centralization or contextual assimilation? Proc. Speech Communication Seminar-74, ed. by G. Fant. (Stockholm: Almqvist and Wiksell), 149-154.

Öhman, S. (1965a) Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. 39, 151-168.

Öhman, S. (1965b) Durations of formant transitions. Speech Transmission Laboratory, QPSR 1/1965, R.I.T., 10-14.

Öhman, S. (1967) Word and sentence intonation: a quantitative model. Speech Transmission Laboratory, QPSR 2-3/1967, R.I.T., 20-54.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllabic nuclei in English. J. Acoust. Soc. Am. 32, 693-703.

Effects of Transition Length on Identification and Discrimination Along a Place Continuum

David Dechovitz† and Roland Mandler†

ABSTRACT

A model of the decision process in discrimination tasks attributes categorical perception along the stop continuum to the transience of the acoustic signal serving as the basis for phonetic divisions [Fujisaki and Kawashima, (1970)]. Two continua of fourteen voiced stop consonants ranging from /ba/ through /ga/ were produced for comparison. In one, transition duration of all three formants was fixed at 30 msec, while in the other, F_2 and F_3 transitions were lengthened to 135 msec, with other parameters left unaltered. Forced-choice identification and same-different discrimination tests were constructed for each set. Responses given by 20 University of Connecticut undergraduates were nearly identical for the two stimulus sets. These results indicate that transience is not the basis for categorical perception of stop consonants.

INTRODUCTION

Research with some classes of synthetically generated speech sounds such as stop consonants has shown that they are perceived in a categorical mode: listeners tend to discriminate between two acoustically different stop consonants to the extent that they can absolutely identify the two as different (Liberman, Harris, Hoffman and Griffith, 1957; Mattingly, Liberman, Halwes and Syrdal, 1971; Pisoni, 1971). In contrast, other classes of speech sounds, such as steady-state vowels, have been found to be perceived in a more nearly continuous mode; listeners are able to discriminate intraphonemic variations (Fry, Abramson, Eimas and Liberman, 1962; Stevens, Öhman and Liberman, 1963; Stevens, Öhman, Studdert-Kennedy and Liberman, 1964).

Differences between consonants and vowels have also been revealed in several recent studies dealing with immediate recall (Crowder, 1971, 1973a, 1973b). There is evidence that a recency effect is characteristic of ordered recall for lists of synthetic stop-vowel syllables contrasting only in their vowel portions. This effect is absent if the syllables contrast only in their stop consonant portions (Crowder, 1971). The parallel between the perceptual differences and the differences in serial recall for stop consonants and vowels has been noted in previous reviews (Crowder, 1971, 1973a, 1973b; Liberman, Mattingly and Turvey, 1972).

---

†Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

Two explanations have been advanced to account for the observed differences between consonants and vowels. The first, termed the "encoding" hypothesis, has its foundations in the assumption that phonetic segments, in the process of articulation, are restructured acoustically so that cues to phonetic identity are distributed over roughly the length of a syllable. However, not all segments undergo restructuring to the same degree, stop consonants being most highly "encoded," and steady-state vowels remaining relatively unencoded (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Liberman, 1970). It is proposed, then, that the complexity of the relation between stop segment and acoustic syllable is more than may be decoded without the use of a special speech processor to strip away the auditory signal and extract the phonetic properties of a syllable (Liberman, Mattingly and Turvey, 1972). Thus, there is for stop consonants no auditory precategorical form available to consciousness. The relatively unencoded vowels, in contrast, may be perceived differently as the auditory characteristics of the signal may be preserved for some short duration.

These inferences are supported by experimental results that point more directly to a special mode of perception for speech. For example, there is evidence that any notion of fixed regions of auditory sensitivity cannot adequately account for the categorical division of the /ba, da, ga/ continuum. Typically, formant patterns controlling consonant assignments are perceived continuously when removed from context and presented for discrimination (Mattingly, et al., 1971; Popper, 1972).

An alternative account, which has been called the cue-duration hypothesis (Pisoni, 1973), asserts that categorical perception is related to the degree to which auditory and phonetic information can be employed in the decision process during a discrimination trial (Fujisaki and Kawashima, 1968, 1969, 1970). The transience of the acoustic signal that serves as the basis for phonetic categories is assumed to affect the persistence of the auditory trace. Thus, stop consonants, for which the acoustic cue is a rapidly changing spectrum (Liberman, Delattre, Gerstman and Cooper, 1954), are poorly preserved in auditory memory. In contrast, synthetic steady-state vowel stimuli, cued by uniform formant frequencies of relatively long duration (Delattre, Liberman, Cooper and Gerstman, 1952) show greater durability.

Two sorts of experimental procedures have provided evidence regarding this hypothesis. The first makes an assessment of the temporal course of recognition memory for vowel and consonant stimuli. In an A-X delayed comparison recognition paradigm, between-category performance was high and independent of delay interval for both consonants and vowels; within-category performance was low and independent of delay interval for consonants, but high for vowels, declining systematically as delay interval increased (Pisoni, 1973).

Manipulations of the acoustic stability of vowel segments have also permitted evaluation of the cue-duration hypothesis. There are reports that vowels are perceived more categorically if their duration and acoustic stability are reduced by placing them in consonant vowel consonant (CVC) syllables (Stevens, 1968; Sachs, 1969; Fujisaki and Kawashima, 1970). However, to the extent that rapidly articulated vowels are substantially

restructured in the sound stream (Liberman, et al., 1967), these results might be assumed to reflect the operation of a speech decoder.

In addition, immediate recall data have revealed no difference between transient diphthongs and steady-state vowels of the same overall duration, either in recency effect or in magnitude of the suffix effect. In fact, the diphthongs showed slightly larger effects, though the difference was not significant (Darwin and Baddeley, 1974). Thus, formant movement seems insufficient to preclude preservation of sounds in auditory memory.

Manipulations of vowel duration alone have sometimes found that vowels presented in isolation in an ABX format are perceived more categorically when they are short (50 msec) than when they are long (300 msec) (Fujisaki and Kawashima, 1970; Pisoni, 1971). However, short vowels of 50 msec duration behaved almost identically to long (300 msec) vowels in the delayed comparison task of Pisoni (1973); the delay interval affected within-category comparisons for both vowel durations in the same way. Further, more recent results from an ABX design failed to show perceptual differences between short (50 msec) and long (300 msec) vowels under certain conditions: differences in discrimination related to stimulus duration were revealed only in one-step comparisons (Pisoni, 1975).

In all these studies, modifications of stimulus transience were made on vowel segments. Stop consonants have not been manipulated similarly, because an early experiment with synthetic speech demonstrated that, as formant transitions are lengthened, stop consonants are converted into semi-vowels (Liberman, Delattre, Gerstman and Cooper, 1956). However, a recent study has demonstrated that the effect of transition duration on stop consonant voicing perception can be primarily attributed to the first formant (Lisker, Liberman, Erickson, Dechovitz and Mandler, in press). And subsequent informal experiments have indicated that manner of articulation is also primarily conveyed by the duration of first formant transition alone: stop perception is preserved when the F_1 transition does not exceed 35 msec, even if the F_2 and F_3 transitions are stretched to nine times that duration.[1] The direction and extent of second formant transition, important for the perceived distinctions among /b, d, and g/ (Liberman, et al., 1954), are apparently effective cues even when that transition is markedly lengthened.

There is, thus, an additional experimental manipulation with which to examine accounts of categorical perception. For if the place distinctions among stop consonants with transitions of natural duration are perceived categorically because the cue-bearing formants are too transient to persist in auditory memory, it would be predicted that stop consonants with transitions 3-4 times their natural length would be perceived more continuously. This prediction was tested in the present experiment.

---

[1]Cragg, R. and Borowitz, L. (1975) Effects of transition length on stop consonant perception. (unpublished).

METHOD

Materials

Two sets of voiced stop consonant-vowel (CV) stimuli were synthesized on the Haskins Laboratories' parallel resonance synthesizer. Each set consisted of fourteen three-formant syllables, all 360 msec duration. The steady-state reached in each syllable was appropriate for an American English /a/, with the first three formants fixed at 769, 1312 and 2861, respectively. One set was a continuum of voiced stops ranging from /ba/ through /da/ to /ga/, with formant transitions of 30 msec. The variable was the pair of starting frequencies of the second and third formant transitions. Stimulus 1 had second and third formant frequencies beginning at 921 and 2525 Hz respectively. For successive stimuli in the series, the F_2 frequencies increased in steps of approximately 75 Hz from 921 to 1920. The starting points of F_3 were placed at (2)2525 Hz, (3)2694 Hz, (4)2964 Hz, (5)2862 Hz, (6)2862 Hz, (7)3026 Hz, (8)3026 Hz, (9)2862 Hz, (10)2862 Hz, (11)2694 Hz, (12)2694 Hz, (13)2525 Hz and (14)2525 Hz. The second set of stimuli was a continuum identical to the first, except that the F_2 and F_3 transitions, while retaining the same beginning and end points, were lengthened in a linear fashion to 135 msec.

Discrimination and identification tests were produced under computer control for each stimulus set. The format of discrimination tests has been the subject of much debate. Fujisaki and Kawashima (1970) propose that ABX and oddball-type discrimination tasks place a heavy tax on short-term memory, thus contributing to, if not inducing categorical discrimination. Similar arguments have been made and supported experimentally by Pisoni (1971). In order to limit the short-term memory load in the present study, and thus decrease the contribution of memory to any categorical effects, the discrimination tests followed a paired same-different paradigm.

Each 70-item identification test was a randomization of the fourteen-item continuum, each stimulus occurring five times. The stimuli were recorded singly, with a three-second interval between presentations.

To test discrimination with a same-different procedure, the stimuli were arranged in pairs, each of which consisted of an A stimulus and a B stimulus that were or were not identical. The A and B stimuli of each pair were determined by (1) pairing each stimulus with the stimulus lying two steps to its right in the continuum, each such pair occurring four times in each of the two possible permutations (A-B, B-A), and by (2) pairing each stimulus with itself one-half the number of times it occurred in pairs with other stimuli. Thus, each discrimination test consisted of 208 stimulus pairs. Within each stimulus pair, items were separated by a 500 msec interval, while successive pairs were separated by a 3 sec interval.

Procedure

The experimental tapes were reproduced on a Crown Series 800 tape deck and presented binaurally through matched and calibrated Telephonics headphones. Subjects for each condition were run in groups of five, each group being presented with the discrimination and identification test sequences for

one of the two synthetic continua. The discrimination test was administered before the identification test to all groups, both sequences being presented in a single experimental session. Before the discrimination test, subjects were told that in each trial, they would hear two stimuli separated by a constant interval, and that their task was to determine whether or not the two were identical. Subjects were instructed to make their judgments on the basis of any audible cues, and to guess if uncertain. Before testing, the experimenter provided each group exposure to ten items of the appropriate sequence, as a brief training session.

In preparation for the identification test, each group of subjects was told that the stimuli would be presented individually, and that they were required to identify each stimulus as belonging to one of three categories (that is, /ba/, /da/, or /ga/).

Subjects

Twenty undergraduates were obtained from the psychology department's subject pool at the University of Connecticut, each receiving one hour of laboratory credit for the experimental session. All subjects had normal hearing and lacked previous exposure to synthetic speech.

Results

The average identification function for each of the two stimulus sets is shown in Figure 1. Each point is based on fifty judgments summed over ten subjects in each condition. Inspection of these figures reveals that subjects partitioned each stimulus continuum into three distinct phonetic categories. These phoneme labeling curves are quite sharp and consistent, and extremely similar in both position and form.

The average discrimination functions for both stimulus conditions are displayed in Figure 2. One sees peaks at the phonetic boundaries and troughs within phonetic categories. The differences between the two functions are not significant by the Mann-Whitney U test [$U(12,12) = 81$, $p < .01$].

We may obtain a fuller view of these results by comparing the obtained discrimination functions to those predicted by an idealized model of categorical perception. Previous workers have developed procedures for predicting from labeling data the level of discrimination to be expected if subjects are able to discriminate only to the extent that they can place different stimuli consistently in different phonemic categories (Liberman et al., 1957; Studdert-Kennedy, Liberman, Harris and Cooper, 1970; Pollack and Pisoni, 1971). For a variety of consonantal discriminations, it has been found that the discrimination functions obtained experimentally lie quite close to the functions predicted on the basis of labeling data. We assume that the difference between the obtained and predicted discrimination functions for a given condition represents a measure of the degree to which that particular condition deviates from the prediction of the idealized categorical perception model. The predicted and obtained discrimination functions are compared in Figures 3a and 3b. Computation of Kendall's tau revealed no significant difference in the form of the predicted and obtained curves for either the

FIGURE 1

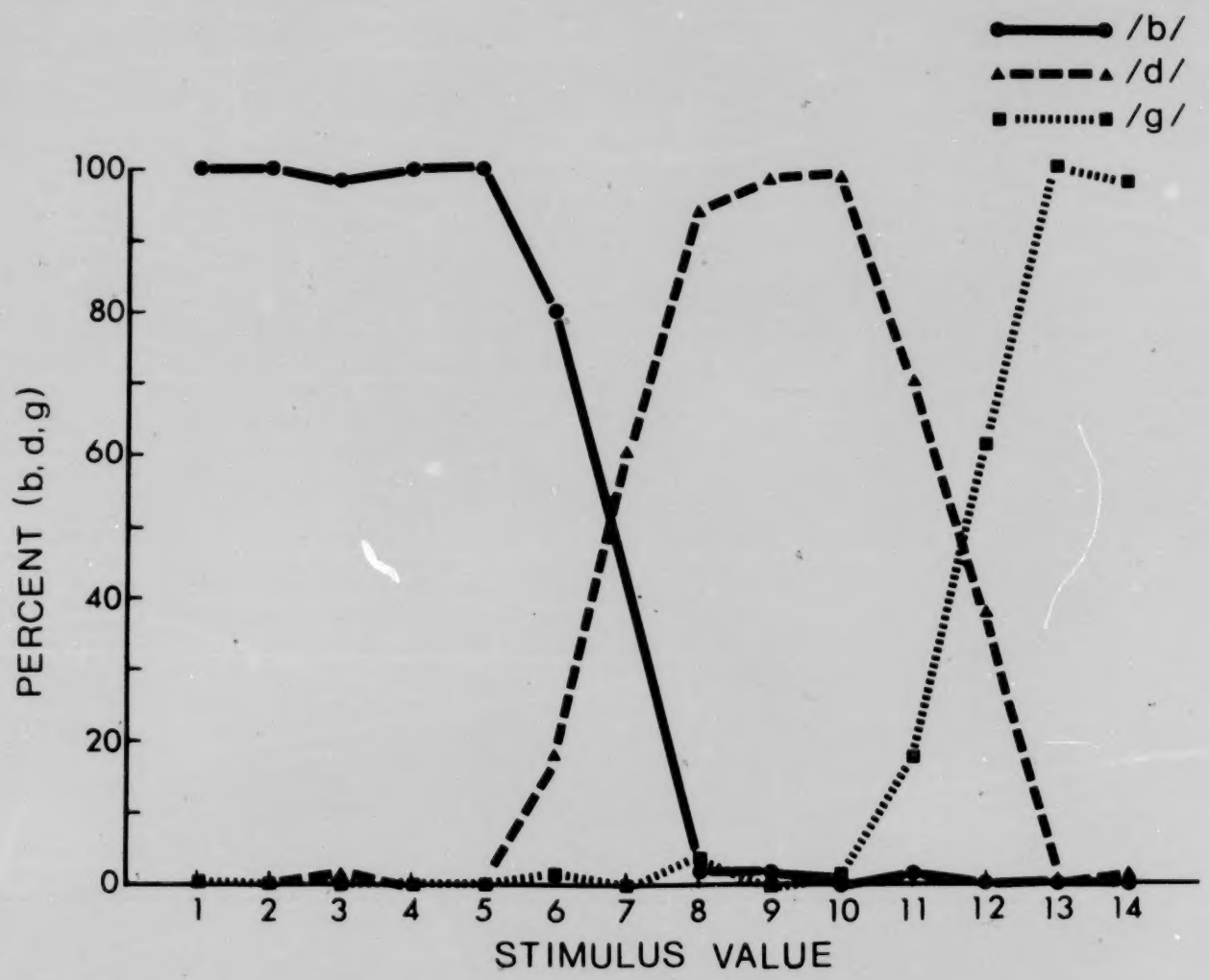

Figure 1: (a) Identification functions for three formant stop consonants with transitions of 30-msec.

(b) Identification functions for three formant stop consonants with 30-msec of transition in F_1, and 135-msec of transition in F_2 and F_3.

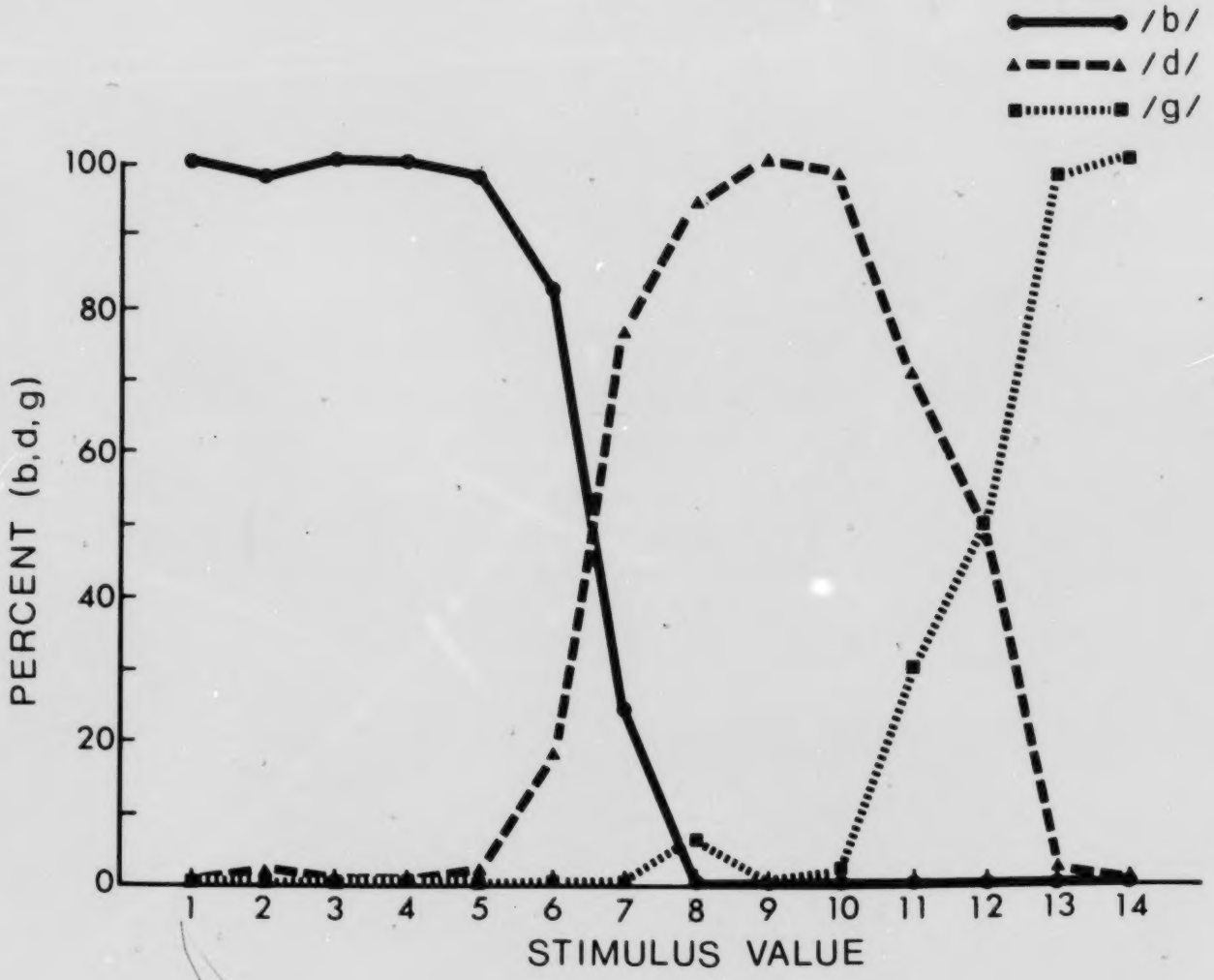

FIGURE 2

Figure 2: Same-different discrimination functions and short-transitioned (30-msec) stop consonants and stop consonants with lengthened (135-msec) upper formant transitions.

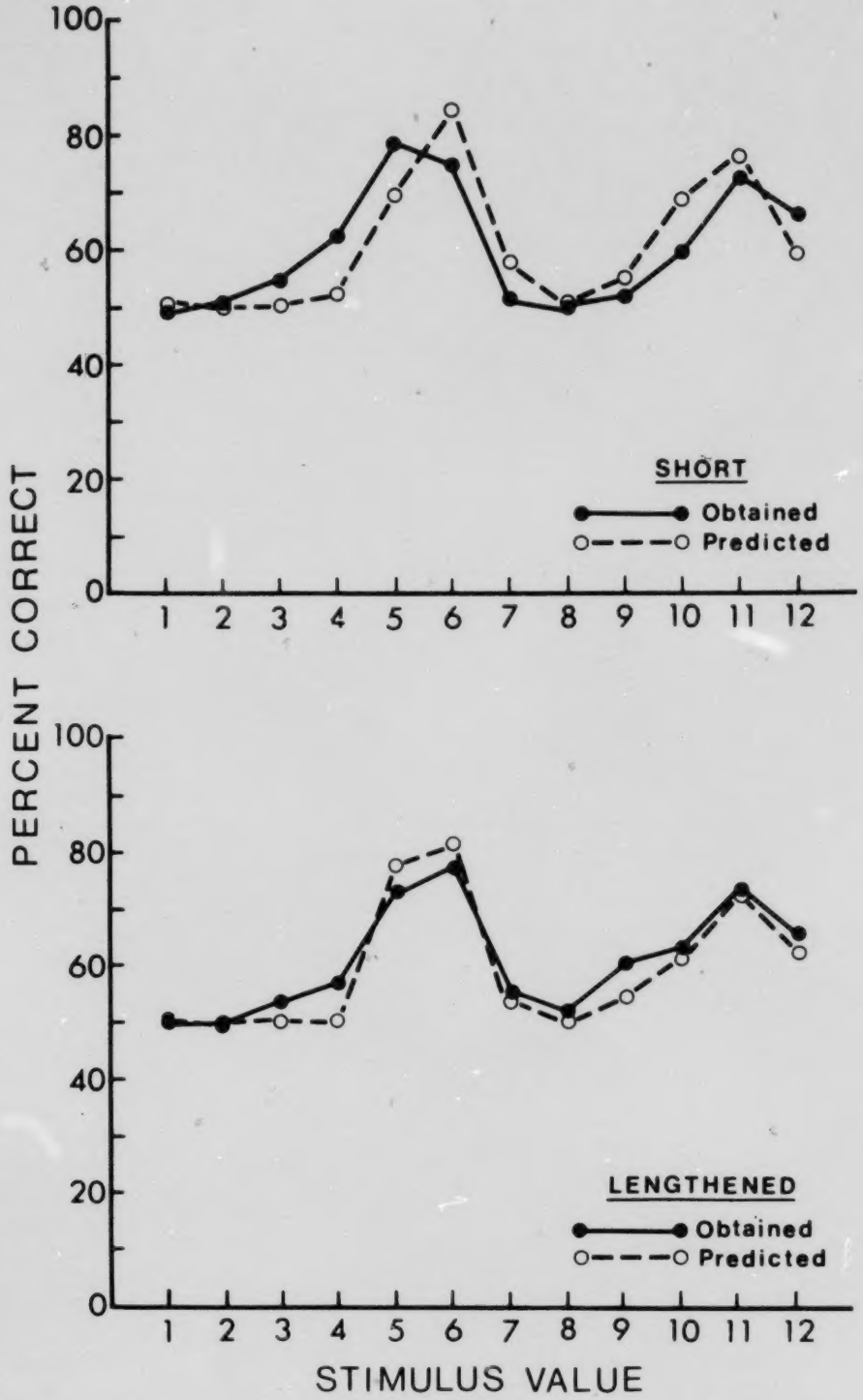

Figure 3: Obtained and predicted (from Pollack and Pisoni, 1971) same-different discrimination functions for short-transitioned (30-msec) stop consonants and stop consonants with lengthened (135 msec) upper formant transitions.

126

short- or long-transitioned consonants (τ=.58, $p < .005$; τ =.84, $p < .001$).

Each point on the discrimination curve represents the sum of percent correct judgments for four presentations: AA, AB, BA and BB. This scoring method separates the relative contributions of discrimination accuracy and possible response bias to observed judgments, since the total for each point is reduced if subjects tend to prefer either same or different judgments. Thus, points for stimulus comparisons 1-3, 2-4, 11-13 and 12-14 are composed of 200 judgments, while the remaining comparisons represent 240 judgments.

In order that the obtained discrimination levels may be compared with d' scores computed for other studies using different discrimination paradigms, the percentages displayed in Figure 2 were transformed to d' scores. False alarm rates were obtained from trials on which subjects responded "different" when the pairs of stimuli were the same, that is, P(DS). Average d' scores for each subject were calculated for stimulus comparisons along the b-d region (pairs 1-6) and for those along the d-g region (pairs 7-12). These values for both stimulus sets are displayed in Table 1. Better discrimination accuracy is indicated by higher d' levels.

TABLE 1: d' values for b-d and d-g regions of continua of short-transitioned (30-msec) and lengthened-transitioned (135 msec) stops.

	d'	
	b-d	d-g
Short	.85	.58
Lengthened	.91	.89

The d' scores were analyzed by means of a two-factor analysis of variance. The main effects of stimulus condition (short vs. lengthened transitions) and region of continuum (b-d vs. d-g) were not significant [$F(1,36)=1.89$; $F(1,36)=3.02$] and there was no significant interaction [$F(1,36)=1.51$].

Though inspection of Figure 2 suggests that discrimination accuracy is depressed along the d-g region of the short-transitioned continuum, examination of average P(D|D) scores (see Table 2) reveals that the depression is the result of unusually high false alarm rates. P(D|D) scores are, in fact, extremely similar along both regions of that continuum.

TABLE 2: P(D|D) values for b-d and d-g regions of short-transitioned stop continuum.

	P(D|D)	
	b-d	d-g
Short	.33	.33

CONCLUSION

The present study indicates that categorical perception cannot be attributed to the transience of acoustic cues underlying phonetic classification; identification and discrimination functions for stop consonant continua with transitions of natural durations (35 msec) and of 3-4 times natural duration (135 msec) are nearly identical.

The role of auditory memory in the discrimination of vowels and stop consonants has received considerable experimental .support (Fujisaki and Kawashima, 1968, 1969, 1970; Pisoni, 1971, 1973, 1975). Our findings do not dispute this role, but do indicate that transience is not a necessary condition of poor auditory storage.

REFERENCES

Crowder, R. G. (1971) The sound of vowels and consonants in immediate memory. J. Verbal Learn. Verbal Behav. 10, 587-596.

Crowder, R. G. (1973a) Representation of speech sounds in precategorical acoustic storage. J. Exp. Psychol. 98, 14-24.

Crowder, R. G. (1973b) Precategorical acoustic storage for vowels of short and long duration. Percep. Psychophys. 13, 502-506.

Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. 6, 41-60.

Delattre, P. C., A. M. Liberman, F. S. Cooper and L. J. Gerstman. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. Word 8, 195-210.

Fry, D. B., A. S. Abramson, P. D. Eimas and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.

Fujisaki, H. and T. Kawashima. (1968) The influence of various factors on the identification and discrimination of synthetic speech sounds. Reports of the Sixth International Congress on Acoustics, Tokyo, Japan, August.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Research Engineering Institute 28, 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the

Engineering Research Institute 29, 207-214.
Liberman, A. M. (1970) The grammars of speech and language. Cog. Psych. 1, 301-323.
Liberman, A. M., F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Liberman, A. M., P. C. Delattre, F. S. Cooper and L. J. Gerstman. (1954) The role of consonantal-vowel transitions on the perception of stop and nasal consonants. Psychol. Monogr. 68, 1-13.
Liberman, A. M., P. C. Delattre, L. J. Gerstman and F. S. Cooper. (1956) Tempo of frequency changes as a cue for distinguishing classes of speech sounds. J. Exp. Psychol. 52, 2, 127-137.
Liberman, A. M., K. S. Harris, H. S. Hoffman and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.
Liberman, A. M. , I. G. Mattingly and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Winston).
Lisker, L., A. M. Liberman, D. M. Erickson, D. R. Dechovitz and R. Mandler. (in press) On pushing the voice-onset-time (VOT) boundary about. Lang. Speech.
Mattingly, I. G., A. M. Liberman, A. K. Syrdal and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.
Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Haskins Laboratories Status Report on Speech Research SR-27, 209-210.
Pisoni, D. B. (1973) Auditory and phonetic codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.
Pisoni, D. B. (1975) Auditory short term memory and vowel perception. Mem. Cog. 3(1), 7-18.
Pollack, I. and D. B. Pisoni. (1971) On the comparison between identification and discrimination tests in speech perception. Psychon. Sci. 24, 299-300.
Popper, R. D. (1972) Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. J. Psycholing. Res. 1, 205-219.
Sachs, R. M. (1969) Vowel identification and discrimination in isolation vs. word context. Quarterly Progress Report no. 93. (Cambridge: M.I.T. Research Laboratory of Electronics), 220-229.
Stevens, K. N. (1968) On the relations between speech movements and speech perception. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 21, 102-106.
Stevens, K. N., S. E. G. Öhman, M. Studdert-Kennedy and A. M. Liberman. (1964) Cross-linguistic study of vowel perception. J. Acoust. Soc. Am. 36, (A)1989.
Stevens, K. N., S. E. G. Öhman and A. M. Liberman. (1963) Identification and discrimination of rounded and unrounded vowels. J. Acoust. Soc. Am. 35, (A)1900.
Studdert-Kennedy, M., A. M. Liberman, K. S. Harris and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Perceptual Integration and Differentiation of Spectral Information Across Intervocalic Stop Closure Intervals

Bruno H. Repp

ABSTRACT

Perceptual interactions between implosive and explosive transitions of intervocalic stop consonants were investigated by preceding stimuli from a synthetic /bɛ/-/dɛ/ continuum by either /ab/ or /ad/. At a very short closure (interstimulus) interval (15 msec), a single stop consonant is heard, with its perceived place of articulation determined primarily by the CV portion of the composite stimulus. However, the vowel consonant (VC) portion exerts a significant positive bias on perception. The strength of this bias varies with the acoustic structure of the VC precursor, which indicates that the bias is due to auditory temporal integration across the closure period. At a longer closure interval (140 msec), two stop consonants are heard if the VC and CV portions convey different places of articulation; a single stop consonant, otherwise. Here a contrast effect is obtained: for example, an /ab/ precursor shifts the /bɛ/-/dɛ/ boundary towards /dɛ/, increasing the probability of hearing /ab-dɛ/. This contrast effect is equally pronounced in the backward direction, when stimuli from an /ab/-/ad/ continuum are followed by either /bɛ/ or /dɛ/. This effect presumably arises at the phonetic level, where differentiation of the speech signal precedes integration.

INTRODUCTION

The present experiment supplements and extends Experiment III of Repp (1977). The earlier experiment investigated whether there are perceptual interactions in the perception of the spectral information preceding and following the closure period of an intervocalic stop consonant (that is, of the implosive and explosive formant transitions), or whether this temporally separated information is perceived independently. Synthetic stimuli from a /bɛ/-/dɛ/ continuum were preceded by either /ab/ or /ad/. When the closure period separating the two stimulus portions (VC and CV) was very short (25 msec), a single consonant was perceived, with its place of articulation determined primarily by the CV portion (the explosive transitions). However, the VC precursors did exert a significant positive bias on perception: the /b/-/d/ phoneme boundary was shifted away from the precursor, relative to a control condition in which only the CV portions were identified. At a very

---

Acknowledgment: This research was supported by NICHD Grant HD01994. Thanks are due to Georgann Witte for her help in analyzing the data.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

long closure interval (265 msec) that permitted the listeners to hear two stop consonants, the VC precursors had no significant differential effect on the perception of the CV portion.

In two other conditions of the experiment, stimuli from an /ab/-/ad/ continuum were followed by either /bɛ/ or /dɛ/, in order to investigate whether perceptual interactions exist in a backward direction as well. At a closure period (115 msec) thought to be sufficient to hear two stop consonants when the VC and CV portions conveyed different places of articulation, there was no differential effect of the two CV postcursors on perception of the VC portion, although the mere presence of a postcursor biased perception towards /b/, relative to a control condition where only the VC portions were identified. In other words, perception of /ad-bɛ/ (or /abɛ/) with the /bɛ/ postcursor was just as likely as perception of /adɛ/ (or /ab-dɛ/) with the /dɛ/ postcursor. At a very long closure period (265 msec), however, there was a small but significant contrastive effect, such that the /ab/-/ad/ phoneme boundary was shifted towards the postcursor, relative to the control condition.

The assimilative perceptual interaction of implosive and explosive formant transitions at very short closure periods was interpreted as evidence of a form of auditory temporal integration, that is, perceptual integration of auditory information before phonetic categorization. That the effect was not merely due to occasional perceptual dominance of the implosive transitions was suggested by separate analyses of the subjects' rating responses within each response category. However, it is theoretically conceivable that the implosive transitions were covertly categorized and influenced the perception of the explosive transitions at the phonetic level, before the temporary category was lost from memory or integrated with the perceptually more dominant category derived from the explosive transitions. In order to investigate this possibility, two acoustically different versions of each precursor were used in the present experiment, one of which was closer to the phoneme boundary on the synthetic VC continuum than the other. If implicit categorization is involved, these within-category acoustic differences should have little effect. However, the auditory integration hypothesis predicts that the acoustically more extreme /ab/ (/ad/) will exert a stronger positive bias than the less extreme precursor from the same category.

A second condition of the present experiment replicated the earlier condition with a 115-msec closure period that had not yielded any differential postcursor effect. Again, stimuli from an /ab/-/ad/ continuum were followed by either /bɛ/ or /dɛ/, but the closure duration was increased to 140 msec. Informal evidence suggested that the 115-msec closure interval perhaps had been too short, so that the task was too difficult for most listeners. It was expected that the 140-msec closure interval would enable the subjects to perceive two different stop consonants without difficulty whenever the implosive and explosive formant transitions specified different places of articulation. A similar condition was included in which stimuli from a /bɛ/-/dɛ/ continuum were preceded by either /ab/ or /ad/. The latter condition investigated forward interaction (the influence of the VC precursor on the perception of the CV portion), while the former investigated backward interaction (the effect of the CV postcursor on the perception of the VC portion). If there was any interaction at all, it was expected to be contrastive,

because it would presumably take place at the phonetic (categorical) level, since the closure interval was long enough to permit categorization of the implosive transitions. It was of interest whether forward and backward interaction conditions would differ in the magnitudes of the interaction effects obtained. Intuitively, forward influences seemed more likely than backward influences; however, the earlier experiment showed the opposite at very long closure durations (265 msec), although the effects were very small.

METHOD

For details of stimulus construction and procedure, the reader is referred to Repp (1977, Exp. III).

Subjects

Nine paid volunteers who had not been subjects in the previous experiment participated.

Stimuli

The stimuli were the same as in the earlier experiment, except for the changes noted below. The first stimulus series was a duplicate of the isolated CV series of Experiment V. It was followed by a series of 180 stimuli grouped into three blocks of 60. Each block contained one randomization of the stimuli from the 7-member CV continuum (with the replication structure described in Repp, 1977) preceded by each of four VC precursors. The closure period was 15 msec. The four precursors were stimuli 1, 3, 6, and 8 from the 7-member VC continuum, stimulus 8 being an additional syllable especially constructed for this purpose and having even more extreme formant transition offset frequencies than stimulus 7. (It was assumed that stimulus 8 would have been consistently identified as /ad/, given that stimuli 6 and 7 were so identified.) The next stimulus series was a duplicate of the isolated VC series of Experiment V. It was followed by a series analogous to the earlier 115-msec series (where the stimuli from the VC continuum were followed by either /bɛ/ or /dɛ/), except that the closure duration was extended to 140 msec. Finally, another series with a 140-msec closure duration was recorded, in which the stimuli from the CV continuum were preceded by either /ab/ or /ad/.

Procedure

All subjects listened to the conditions in the order described above. The initial CV series and the following 15-msec series were repeated once. (Two subjects omitted the repetition of the CV series.) The instructions were the same as in Experiment V. A 6-point rating scale was used except in the two 140-msec series, where the task was to indicate whether one or two medial consonants were heard.

RESULTS

The results of the 15-msec condition are shown in Figure 1a. The dashed line represents the results for CV syllables in isolation. The function was somewhat less steep than in the earlier experiment, but this was mainly due to

the subjects' use of less extreme ratings; the two phoneme categories were well separated. The precursor functions looked somewhat different from those in the earlier study. In particular, the precursors seemed to have no differential effect at the /b/-end of the CV continuum, except for shifting the ratings slightly towards /b/. Towards the /d/-end of the continuum, however, a strong assimilative precursor effect emerged, in agreement with the earlier results. In addition, those precursors that were closer to the phoneme boundary produced a weaker effect than the precursors from the ends of the VC continuum, as predicted under the auditory integration hypothesis.

The statistical analysis showed a significant between-category effect of VC precursors (precursors 1 and 3 vs. 6 and 8: $F_{1,8} = 8.9$, $p < .025$), an interaction of this effect with position on the CV continuum ($F_{4,32} = 4.0$, $p < .025$; this significance level was relatively low only because of high between-subject variability), a significant interaction of the within-category precursor effect (precursors 1 and 6 vs. 3 and 8) with position ($F_{4,32} = 4.6$, $p < .01$), and a significant triple interaction ($F_{4,32} = 5.5$, $p < .01$). The main effect within precursor categories did not reach significance.

In order to clarify these results, weighted average ratings were computed conditionally on the two response categories (ratings 1-3 = B; ratings 4-6 = D). These data are plotted in Figure 1b. Separate analyses of variance were conducted on B and D responses, including (the first or last) four positions on the CV continuum. (These analyses were conducted on unweighted conditional ratings, while Figure 1b represents weighted means; therefore, the statistical results may deviate slightly from the graphical representation.) B responses showed a highly significant effect of position ($F_{3,24} = 43.3$, $p << .01$), obviously due to the increase in ratings between positions 3 and 4. The between-category precursor effect just fell short of significance ($F_{1,8} = 5.2$, $p < .06$), indicating a tendency in the expected direction. All other effects were far from significant. The D responses showed significant effects of position ($F_{3,24} = 3.9$, $p < .025$)--Figure 1b shows the effect to be very small--between precursor categories ($F_{1,8} = 15.6$, $p < .01$) and within precursor categories ($F_{1,8} = 6.0$, $p < .05$). The between-category effect also interacted with position ($F_{3,24} = 3.5$, $p < .05$), but this effect was negligible.

The results for D responses support the prediction that within-category acoustic differences between VC precursors would affect the degree of the assimilative precursor effect. Such a gradual effect most likely reflects auditory integration of implosive and explosive transitions; implicit categorization of the implosive transitions seems unlikely. It should be noted that the VC precursors were very consistently identified in isolation (cf. Figure 1c): stimuli 1 and 3 received 100 and 98 percent B responses, respectively; stimuli 6 and 7 received 98 and 100 percent D responses, respectively, so that it was safe to assume that stimulus 8 (the actual precursor) would have received 100 percent D responses, too. The between-category precursor effect was somewhat larger than the within-category effect, but this probably reflected the fact that the acoustic difference between categories was larger than that within categories.

The results for the 140-msec VC (backward interaction) series are shown in Figure 1c. The labeling function for VC stimuli in isolation and the

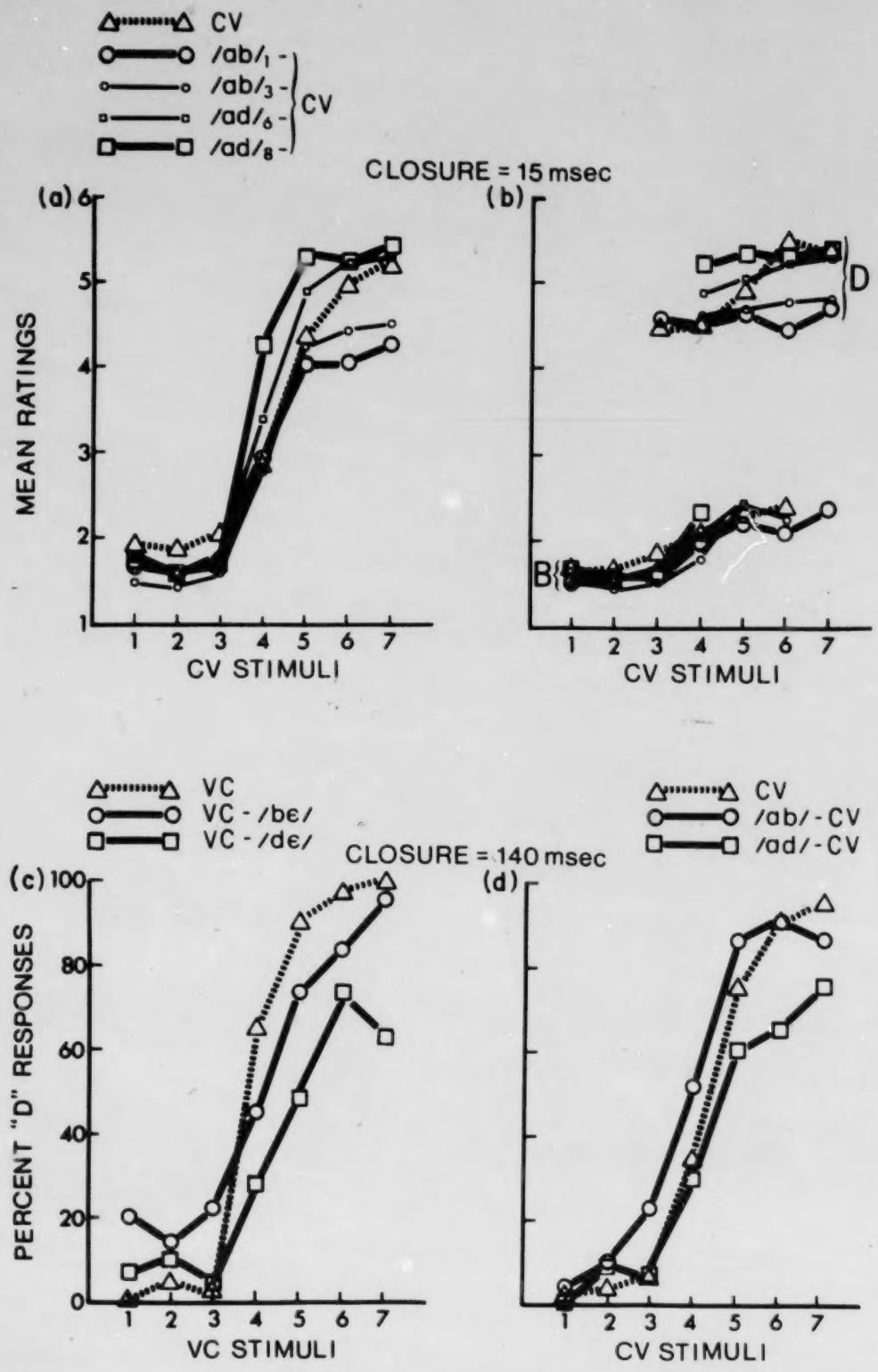

Figure 1:
(a) Mean "D-ness" ratings of syllables from a /bɛ/-/dɛ/ continuum in isolation (CV) and when preceded by a 15-msec silent interval and one of four VC precursors. The subscripts denote the position of the VC precursors on the /ab/-/ad/ continuum. (b) The same data, analyzed separately for the two response categories (B = ratings 1-3; D = ratings 4-6). (c) Average percentage of "D" responses to syllables from an /ab/-/ad/ continuum in isolation (VC) and when followed by one of two CV postcursors after a 140-msec silent interval. (d) Average percentage of "D" responses to syllables from a /bɛ/-/dɛ/ continuum in isolation (CV) and when preceded by a 140-msec silent interval and one of two VC precursors.

general shape of the postcursor functions replicated the 115-msec condition of Repp (1977, Exp. III, Figure 10a). However, in contrast to the earlier results, there was a clear difference between the two postcursor functions in the predicted direction, viz., a contrast effect ($F_{1,8} = 13.6$, $p < .01$) which did not interact with position on the VC continuum. Seven of the nine subjects showed a contrast effect, one showed no clear effect, and one a slight assimilative effect. Thus, the increase in closure duration from 115 to 140 msec apparently facilitated the task, so that a contrast effect could emerge. At the 115-msec closure duration, some subjects still seemed to have a tendency to integrate implosive and explosive transitions; this tendency was no longer present at the 140-msec closure duration.

The results for the new 140-msec CV (forward interaction) series are shown in Figure 1d. It can be seen that this series, too, showed a pronounced effect of the VC precursors on the perception of the CV portions ($F_{1,8} = 11.5$, $p < .01$). Contrast effects were shown by six subjects; the remaining three showed no clear effects. The precursor functions followed the isolated-CV function more closely than the postcursor functions in Figure 1c followed the isolated-VC function: VC syllables sounded more labial when followed by a CV postcursor, but not vice versa. The former effect may reflect merely a peculiarity of the present stimuli, but it may also indicate that implosive transitions are perceptually less stable than explosive transitions.

DISCUSSION

The implosive and explosive formant transitions surrounding the closure period of an intervocalic stop consonant provide an extreme example of temporally distributed information for a single phoneme. In order to perceive this information as a single event, an integrative perceptual process is needed. The present experiment, together with its predecessors (Repp, 1976, 1977) suggests that there are at least two levels at which acoustic cues may be integrated. One level is auditory, the other phonetic in nature; this distinction follows common usage in speech perception research (for example, Studdert-Kennedy, 1976).

Auditory temporal integration is apparent in the situation where conflicting implosive and explosive transitions separated by a very brief closure period lead to the perception of a single phoneme (Dorman, Raphael, Liberman, and Repp, 1975). This "backward masking" effect was replicated here using an identification-judgment task. The analysis of conditional ratings, as well as the within-category precursor effect, specifically support the notion of auditory temporal integration with higher weighting of more recent information; they make an interpretation in terms of true recognition backward masking (that is, interruption of processing--Massaro, 1975) seem unlikely.

The temporal limits of this auditory integration may have just been reached in the inconclusive 115-msec condition of the earlier study. There seemed to be a region of uncertainty beyond the region over which the "backward masking" effect primarily operates. An estimate of the total integration period may be obtained by adding the duration of the spectral information being integrated to the maximal temporal separation. Thus, a value of approximately 200 msec is obtained, which is of the same magnitude as Huggins' (1975) estimate from the perception of temporally segmented speech.

Huggins hypothesized that "the ear tries to integrate into a single percept any two relatively similar events that coexist in echoic storage, and only becomes able to treat them as separate events if they do not coexist in echoic storage" (p. 156). If it is accepted that any implosive and explosive transitions are "relatively similar," the 200-msec auditory integration period obtained here may well reflect the duration of an echoic store. The perceptual predominance of the more recent information (explosive transitions) may be due to the fact that it resides longer in auditory memory (adopting Huggins' analogy of storage with a delay line). There are a number of related findings in the auditory literature suggesting a temporal integration period of about 200 msec (see Huggins, 1974, 1975, for a brief discussion). This may or may not be a coincidence; certainly, stimulus and task factors play a role in determining the precise interval over which integration occurs.

A second process of temporal integration must be postulated to account for the perceptual distinction between single and geminate stop consonants. When implosive and explosive transitions signalling the same place of articulation are separated by less than about 200 msec of silence, they are perceived as a single phoneme (Pickett and Decker, 1960; Repp, 1976). This integrative process, whose "time window" of about 300 msec clearly exceeds that of auditory temporal integration, most likely takes place at the phonetic level. Presumably, a phonemic percept is established on the basis of the implosive transitions plus silence, but if an identical phonemic percept is arrived at within a certain time span (on the basis of explosive transitions occurring 100-200 msec later), a higher-order perceptual rule integrates the two into a single percept. This perceptual rule may well represent the listener's implicit knowledge about articulatory processes, and therefore it may have no parallel outside speech perception. Moreover, Pickett and Decker (1960) have shown its extreme sensitivity to contextual variables, particularly rate of speech. Similar variability of temporal parameters is rarely found in auditory (nonspeech) perception.

The process of phonetic integration is parallelled by the contrast effects observed at the 140-msec closure duration in the present experiment. These contrast effects seem logically prior to integration: integration is possible only if implosive and explosive transitions are perceived as signalling the same place of articulation. However, the nonindependence of the perception of these spectral cues, both forward and backward in time, is in itself evidence of a perceptual process that operates on fairly long temporal segments of the speech signal. What has been called here phonetic integration may be considered merely a consequence of this more general process of "coperception." Coperception--the process that enables the listener to interpret a substantial stretch of the acoustic information in parallel--serves both to integrate and to differentiate acoustic cues. Perceptual contrast effects reflect the tendency of the perceptual system to differentiate; integration follows if differentiation does not occur. The perceptual interactions and individual differences observed by Repp (1977) at closure durations beyond the single-geminate boundary need further corroboration before it is concluded that perceptual interactions between implosive and explosive transitions and the phonetic integration responsible for the single-geminate distinction are not, in fact, due to the same underlying process. Coperception subsumes constraints imposed on perception by a short-lived auditory memory, as well as higher-order phonetic processes that are not

subject to similar limitations. While the former are characterized by a more or less fixed time constant, the latter are better described in terms of flexible perceptual rules. Both processing levels cooperate in differentiating and integrating the temporally distributed information in the speech signal.

REFERENCES

Dorman, M. F., L. J. Raphael, A. M. Liberman and B. H. Repp. (1975) Some maskinglike phenomena in speech perception. Haskins Laboratories Status Report on Speech Research SR-42/43, 265-276.

Huggins, A. W. F. (1974) On perceptual integration of dichotically alternated pulse trains. J. Acoust. Soc. Am. 56, 939-943.

Huggins, A. W. F. (1975) Temporally segmented speech. Percep. Psychophys. 18, 149-157.

Massaro, D. W. (1975) Preperceptual images, processing time, and perceptual units in speech perception. In Understanding Language. An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics, ed. by D. W. Massaro. (New York: Academic), pp. 125-150.

Pickett, J. M., and L. R. Decker. (1960) Time factors in perception of a double consonant. Lang. Speech 3, 11-17.

Repp, B. H. (1976) Perception of implosive transitions in VCV utterances. Haskins Laboratories Status Report on Speech Research SR-48, 209-233.

Repp, B. H. (1977) Perceptual integration and selective attention in speech perception: Further experiments on intervocalic stop consonants. Haskins Laboratories Status Report on Speech Research SR-49, 37-69.

Studdert-Kennedy, M. (1976) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic), pp. 243-293.

Musical Skill and the Categorical Perception of Harmonic Mode*

Mark J. Blechner†

ABSTRACT

Professional musicians and nonprofessional subjects participated in a series of experiments designed to determine whether musical chords are perceived categorically and whether musical skill affects such perception. In a preliminary task requiring identification of prototype minor and major chords, nearly half of the nonprofessional subjects could not perceive the difference between the two chords. This finding suggested a subdivision of the nonprofessional subjects into two groups--nonprofessional, high-skill (NP-H) and nonprofessional, low-skill (NP-L).

The three groups of subjects--professional musicians, NP-H, and NP-L--were then asked to identify and discriminate a set of musical triads in which the central tone (the musical third) varied in discrete steps along a continuum of frequency between the prototypes of minor and major chords. The pattern of results indicated categorical perception for most of the professional musicians and the NP-H subjects, but not for the NP-L subjects. A detailed analysis of order effects in the discrimination task suggested that the NP-H subjects, in certain conditions, have the option of making discriminations based either on the category membership of the chords or solely on their auditory characteristics. The professional musicians, however, appear to use category membership consistently in the discrimination task, even when such a strategy is not optimal.

---

*This paper is based on a dissertation submitted to Yale University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

†Also Yale University, New Haven, Connecticut.

Acknowledgment: The author wishes to thank Ruth Day, his advisor, and Wendell Garner, Irvin Child, Robert Crowder, Alice Healy, James Cutting, Michael Studdert-Kennedy, Robert Abelson, and Andrea Levitt for much helpful advice. This research was supported by National Institute of Mental Health Training Grant PHS5T01MH05276-27 to Yale University, National Institute of Child Health and Human Development Grant HD-01994 to the Haskins Laboratories, and a Dissertation Grant to the author from the Graduate School of Yale University.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 1977]

In a second experiment, the same subjects were asked to identify and discriminate a set of single tones varying along a frequency continuum. These single tones were identical to the central tones of the chords in Experiment I, and thus constituted the "minimal cue" to the minor-major distinction in those chords. In the identification task with single tones, all subjects partitioned the stimulus continuum into binary categories, but the discrimination functions did not indicate categorical perception. The results of both experiments suggest that categorical perception may occur for musical triads when the critical dimension is determined by a complex relationship of three frequencies. Categorical perception does not occur, however, when only the central tone, whose pitch cues the distinction between minor and major, is presented in isolation. These findings closely parallel data from studies of speech perception which show that linguistic experience affects the perception of speech syllables, but not the perception of nonspeech cues isolated from such syllables. The present results also indicate that experience can affect both the level and pattern of performance, and that significant patterns of results may be masked by the procedure of data averaging that has been common in studies of categorical perception.

INTRODUCTION

Experience is often thought to affect perception in a quantitative manner. For example, it can enable the perceiver to process information more quickly or with greater accuracy. However, experience may also affect perception in qualitative ways. For example, when a stimulus can be perceived along several dimensions, experience may alter the ease with which a person can attend to various of its aspects. In such cases, our main interest shifts from concern with the subject's perception of the stimulus as a whole to issues concerning underlying processes. Such issues include: a) the extent to which a subject can perceive individual dimensions; b) whether the level of perception of one dimension facilitates, interferes with, or has no effect on the perception of another dimension; c) the conditions under which various perceptual processes are mandatory or optional; d) the ways in which various experimental operations reflect such phenomena; and e) how the interaction of such component processes results in perception of the stimulus as a whole.

These concerns have become increasingly important in the study of a phenomenon known as categorical perception, which has been studied intensively over the past 25 years. Stimulus dimensions that are perceived categorically are thought to be processed in a manner qualitatively different from the way most stimulus dimensions are processed. For most stimulus dimensions, subjects can discriminate between many more stimuli than they can identify (Miller, 1956). Categorical perception, however, is characterized by a set of psychophysical functions indicating that the ability to discriminate between stimuli is limited by the ability to identify those stimuli into discrete categories.

In order to determine whether a stimulus dimension is perceived categorically, both identification and discrimination tasks are needed. In the identification task, the subject is given an array of stimuli and must classify these stimuli into categories specified by the experimenter. In a discrimination task, the subject need only be able to perceive differences in sound without having to label the sound. For example, in the "oddball" form of the discrimination task, the subject is given three stimuli sequentially and told that two are the same and one is different. The task is then to determine whether the "oddball" is the first, second, or third stimulus.

Categorical perception occurs when the following pattern of results is obtained: a) subjects identify stimuli very reliably into discrete categories with nearly 100 percent consistency at most points on the stimulus continuum and with a sharply demarcated shift in the function between categories, known as the "category boundary;" b) when stimuli to be discriminated are both from the same category, performance is at or near chance, yielding a "trough" in the discrimination function: c) when stimuli are identified as being from different categories, discrimination performance is very high, yielding a "peak" in the discrimination function; d) there is a reasonably good fit between the obtained discrimination function and a mathematically-predicted discrimination function that is derived solely from the probability that a stimulus is identified as belonging to one category or another.

Categorical perception was first observed with an array of speech stimuli (Liberman, Harris, Hoffman and Griffith, 1957). The slope of initial second-formant transitions in consonant-vowel syllables was varied in discrete steps, but identification and discrimination appeared to be determined by the phonemic categorization of the stimuli as either /ba/, /da/, or /ga/. Subsequently, categorical perception was observed for other linguistic dimensions, such as voice-onset time (VOT) (Abramson and Lisker, 1965, 1970; Lisker and Abramson, 1970). However, for many years categorical perception was never observed for nonlinguistic sounds, even when such stimuli seemed to be acoustically related to speech stimuli. Liberman, Harris, Kinney and Lane (1961), for example, found no evidence for categorical perception of synthetic stimuli derived from inverted speech spectrograms. Similarly, second formant transitions, which are sufficient cues for place of articulation in speech syllables, were isolated from their speech context and perception of them was not categorical (Mattingly, Liberman, Syrdal and Halwes, 1971; Miyawaki, Strange, Verbrugge, Liberman, Jenkins and Fujimura, 1975). From these results, it was suggested that categorical perception was unique to speech stimuli, and that it might, in fact, imply the existence of two modes of perception--one for speech and one for nonspeech. This interesting hypothesis was challenged by the subsequent finding of categorical perception for several nonlinguistic sounds: sawtooth waves differing in rise time, which are heard as plucked or bowed violin strings (Cutting and Rosner, 1974); sequences consisting of a burst of noise followed by a buzz, in which the relative onset of the two sounds was varied (Miller, Wier, Pastore, Kelly and Dooling, 1976); and staggered sine waves in which the relative onset of the two tones was varied (Pisoni, in press). In the visual modality, categorical perception has also been found for nonlinguistic stimuli, using light-emitting diodes activated by square waves, in which the period of the wave was varied in 6 or 8

msec increments (Pastore, 1976).

Given these results, several researchers have hypothesized that categorical perception might be due to a basic sensory limitation in the ability to resolve temporal variation (Jones, 1976). Hirsh (1959) had found that the minimum discriminable interval between the onset of two stimuli was 20 msec. This was close to the value of the category boundary for VOT, the noise-buzz stimuli, the staggered sine waves, and the plucked and bowed sounds. The sensory-limitation hypothesis was further supported by the observation of categorical-like perception of stop consonants in infants (Eimas, Siqueland, Jusczyk and Vigorito, 1971; Eimas, 1975) who have had only minimal exposure to speech sounds. These results, by themselves, could suggest that phonetic processors responsible for perceiving in the speech mode are innate in humans; but it has since been found that chinchillas which, of course, never acquire speech, also perceive stop consonants in a categorical fashion (Kuhl and Miller, 1975). Moreover, Jusczyk, Rosner, Cutting, Foard and Smith (1977) have obtained categorical-like results for infants using nonlinguistic pluck-and-bow stimuli. Only the innate sensory-limitation hypothesis can account for all of these results with infants and chinchillas.

Nevertheless, the preponderance of data attributing categorical perception to the resolution of temporal variation might be a by-product of the fact that most studies used stimuli that varied a temporal dimension. One may wonder, then, whether categorical perception could also occur in stimuli that do not vary over time. Experiments with steady-state vowels are relevant to this question, but these data have been ambiguous. One study (Liberman, Harris, Eimas, Lisker and Bastian, 1961) found completely noncategorical perception for vowels, with gradually changing identification functions and essentially flat discrimination functions. Other studies (for example, Stevens, Liberman, Studdert-Kennedy and Öhman, 1969; Pisoni, 1971) have found peaks in discrimination functions for vowels, but within-category discrimination was still significantly above chance; that is, there were no "troughs." From such results, certain authors have begun to consider that categorical perception need not be an all-or-nothing effect. Studdert-Kennedy, Liberman, Harris and Cooper (1970), for example, assert "that there is a difference in the degree to which consonants and vowels are categorically or continuously perceived... (p. 238)."

Fujisaki and Kawashima (1969; 1970) have attempted to account for discrepant findings with vowels by demonstrating that perception of vowels appears to be more categorical as the duration of the stimulus decreases. They therefore conclude, as does Pisoni (1971), that short-term memory may be a potent factor in determining whether a sound is perceived categorically. Nevertheless, such explanations do not rule out the possible importance of phonetic, that is, speech-specific coding, as well.

The question remains, then, whether any steady-state nonlinguistic stimuli yield conclusive evidence of categorical perception. Musical chords are potentially good candidates for study of this issue. They can vary continuously in their physical characteristics, but only a small subset of the possible chords are used in musical practice. For example, in a typical

triadic (three-tone) major chord, the central tone, (known as the "third"), is approximately 5/4 of the frequency of the lowest tone (known as the "root"); in a typical minor chord, the third is approximately 6/5 the frequency of the root. It is possible to substitute for the major or minor third a tone that is, say, midway between the prototypes for minor and major, for example, one that is 49/40 of the root. Such a chord is unlikely to occur in the tonal Western music of the last 200 years, (at least when played correctly), although it is easy to produce.

Several studies suggest that chords are perceived categorically, but none of them has provided all of the relevant tasks or analyses necessary to make this claim. Locke and Kellar (1973), for example, found seemingly discontinuous discrimination functions in some musicians for triadic chords that varied the central tone gradually from minor to major. However, they used an A-X discrimination procedure that required a signal detection analysis of their data, which may have been carried out inappropriately. [For a complete discussion of the problems with their data analysis, see Pastore, (1976).]

The present study, therefore, seeks to determine whether the perception of musical chords does, in fact, fulfill all of the criteria stated above for categorical perception, namely: consistent identification categories with a sharp boundary between them; peaks and troughs in the discrimination function; and a good fit between the obtained and predicted discrimination functions.

INDIVIDUAL DIFFERENCES AND THE ROLE OF EXPERIENCE

Early studies of categorical perception posited two possible explanations for the phenomenon: a) that the ability to discriminate across category boundaries is learned ("acquired distinctiveness" was the term often used) and that the lack of ability to discriminate items from the same stimulus category was also learned ("acquired similarity"; Liberman, Harris, Eimas, Lisker and Bastian, 1961); b) that the ability to perceive distinctions between phoneme categories and the lack of ability to perceive intracategory acoustic differences is innate.

To differentiate between these two theories of categorical perception, two basic research strategies have been used. One is to study the perception of infants, who have not yet learned speech. Eimas and his colleagues, for example, have found (as noted above) that neonates seem to perceive stop consonants in a discontinuous, categorical fashion, which suggests that phonetic perception, or the ability to perceive the acoustic cues underlying phonetic perception, is innate.

Another research strategy is to study native speakers of different languages, using stimulus distinctions that are linguistically relevant in one language but linguistically irrelevant in another language. For example, rounding distinguishes between different vowels in Swedish, but in English the rounding distinction is not phonemic. Yet the identification and discrimination of both rounded and unrounded vowels are nearly identical for both American and Swedish subjects, suggesting that linguistic experience does not affect the perception of vowels (Stevens, Liberman, Studdert-Kennedy and

Öhman, 1969). Lisker and Abramson (1970), on the contrary, found marked differences in the identification boundaries for stop consonants in American, Spanish, and Thai speakers, suggesting that experience does play a prominent role in the perception of these sounds. Correlated changes in the discrimination functions of English and Thai speakers were reported by Abramson and Lisker (1970).

Further convincing evidence of the role of linguistic experience has come from Miyawaki et al. (1975), who studied the perception of the liquids /r/ and /l/, a distinction that is phonemic in English but not in Japanese. Native speakers of English yielded functions suggesting categorical perception, while Japanese speakers did not. In fact, the Japanese speakers could not perceive any difference between the two phonemes. This study also compared the two types of speakers on the perception of another, nonspeech sound derived from the /ra/ and /la/ syllables. These syllables can be distinguished acoustically by the direction of the third formant (F_3) transition: falling in /la/ and rising in /ra/. When this third formant transition, the minimal acoustic cue to the liquid distinction, is presented in isolation, it does not sound like speech. Miyawaki et al. found that this isolated F_3 transition was discriminated equally well by Japanese and American subjects, and that there was no evidence for categorical perception of the transitions in either group. The authors conclude that the absence of categorical perception for the F_3 transitions in either group is evidence that perception of /ra/ and /la/ results from the operation of a speech-specific processor. They acknowledge an alternative interpretation, that the different results with speech and nonspeech sounds are due to auditory processing of an acoustic interaction between the minimal acoustic cue and its acoustic context, but they reject this explanation:

> In all cases, a result that tends to put the effect in the speech mode could, of course, be interpreted alternatively as a purely auditory interaction between the cue and the constant acoustic context to which it is always added in the speech patterns. But such an interpretation is empty unless one can make sense of it in terms of what is known, on other grounds, about auditory perception (p. 335).

Although Miyawaki et al. suggest that the categorical perception of the American subjects is a result of learning, a study with infants suggests an alternative interpretation of their data. Eimas (1975) found that infants perceive liquids categorically. Therefore, it may be that the mechanism underlying categorical perception is innate in both American and Japanese subjects and that learning accounts for the data of the Japanese subjects; that is, the innate ability to perceive the /r/-/l/ distinction may be lost through disuse in Japanese speakers.

In the case of musical chords, there is evidence that experience may be responsible for categorical perception, but the data are not conclusive. For example, the subject population of Locke and Kellar (1973) consisted of graduate music students from the New England Conservatory, obviously with much musical experience, and a random sample of nonmusicians. However, in neither

group were the amount and kind of musical experience carefully specified. Thus, the data point to marked individual differences in discrimination functions, but neither the musicians nor the nonmusicians group produced homogeneous results.

The present study, therefore, attempts to specify more precisely the amount and kind of musical experience that would be relevant to discrimination between musical chords. One might hypothesize, for example, that acoustic factors aside, categorical perception would be most likely when subjects are very accustomed to using the available stimulus categories. Categorical perception would be least likely when intracategory acoustic differences are highly familiar to the subject. Keyboard musicians, for example, can produce complete chords with their instruments, while they do not have immediate control of the intonation of individual tones. Therefore, they may be more likely to perceive chords categorically than, for example, a trombone player or a singer, who can only produce a single note at a time, and who can vary his intonation considerably.

The fact that the stimuli in the present experiment are composed of simple sine waves also determined the procedure of subject selection. A group of keyboard players was recruited whose primary instrument was the organ, since the timbre of that instrument can reasonably approximate simple sine waves. In summary, the goal of the selection procedure was to obtain one group of musical subjects who were as familiar as possible with the type of chord stimuli used in this experiment as most people are familiar with speech stimuli.

EXPERIMENT I - METHOD

Stimuli

The stimuli, chords composed of three sine waves of different frequencies, were constructed from an original set of eleven sine waves. These were generated by a Hewlett Packard Test Oscillator (model 208A) and were accurate to $\pm$ .1 Hz. The individual sine waves were trimmed to 250 msec in duration, and were then digitized and stored on computer disc, using the Pulse Code Modulation (PCM) system at Haskins Laboratories. The individual tones were then combined to form nine three-tone chords. In each chord, the high and low pitches were always 392 and 262 Hz, respectively. The central tone varied in discrete steps of 2.32 Hz, from 311 to 329.6 Hz. (The complete set of stimuli is displayed graphically in Figure 1.) The chords were labelled #1 - 9, corresponding to the lowest through the highest central tones. Those at each end of the continuum, chords #1 and #9, were designated as the prototypes for minor and major chords, respectively. Chords were stored on computer disc in digital form until the time of tape recording, when they were reconverted to analog form.

Tapes

Tapes were recorded using the PCM system. A number of tapes were prepared in order to familiarize the subjects with the stimuli and to test

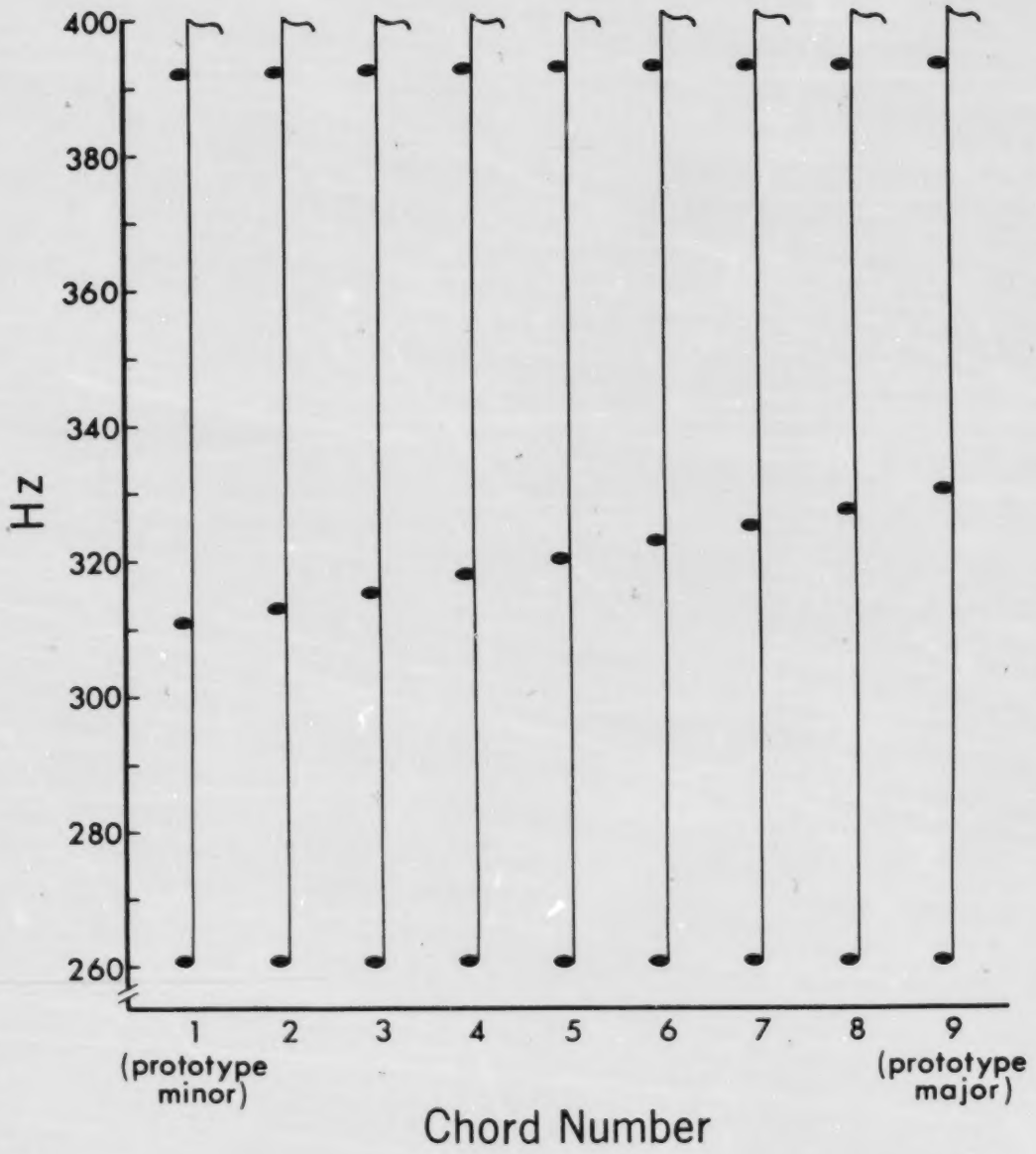

Figure 1: Schematic representation of chords used in Experiment I.

146

their identification and discrimination. A prototype display tape was recorded which contained several examples of the prototype minor and major chords. In this tape, a sequence of three minor chords and three major chords was followed by a pause; then came a sequence of two minor chords and two major chords also followed by a pause; and finally one token of each type of chord.

The prototype test tape consisted of 30 tokens of each of the minor and major prototypes in random order. The interstimulus interval (ISI) was 2 sec.

The identification test tapes consisted of two blocks of 72 chords, with an ISI of 3 sec. Each block contained eight tokens of the nine possible chords in random order. The identification practice tape consisted of one complete set of the nine possible chords in random order, again with a 3 sec ISI.

The oddity-discrimination test tape consisted of two blocks of 90 triplets of chords. The chords in each triplet were separated by an ISI of 500 msec, and an interval of 5 sec separated the offset of each triplet from the beginning of the next triplet. Two of the chords in each triplet were identical, and one of them differed from the other two by one or two steps on the stimulus continuum. There were eight possible one-step discriminations (chords #1-2, 2-3, 3-4, ...8-9) and seven possible two-step discriminations (chords #1-3, 2-4, 3-5, ...7-9). For each pair of stimuli (A-B), all six possible permutations of the stimuli (that is, AAB, ABA, BAA, ABB, BAB, and BBA) were included in each block. The oddity-discrimination practice tape consisted of ten randomly chosen triplets of chords that obeyed the same temporal constraints as the test tape.

Subjects

Five organists who were graduate students in the Yale School of Music constituted the Professional Musicians group. These subjects all spent the bulk of their working time practicing their instrument, an average of 28.6 hours per week. They were paid $2.00 per hour for participating in the experiment.

It was originally planned to have two groups of subjects--professional musicians and nonmusicians. The latter group consisted of undergraduates from the Introductory Psychology course at Yale University, who received course credit for their participation in the experiment. However, the sign-up sheet for these subjects asked for "nonmusicians," which apparently was understood by some subjects to mean people who were not music majors, and by others as meaning people who had no musical training at all. The eleven undergraduates who were thus recruited for the experiment differed markedly in musical ability. Their data readily suggested a post hoc subdivision into two groups: six "nonprofessionals, highly-skilled" (NP-H), and five "nonprofessionals, low-skilled" (NP-L). (The details and rationale of this subdivision are discussed below in the Results section.) All of the subjects reported no history of hearing trouble.

Apparatus

The tapes were played on an Ampex AG-500 tape recorder, and the stimuli were presented through calibrated Telephonics headphones (Model TDH39-300Z), at an intensity level of 68 dB SPL.

In the identification tasks, in which there were always two response alternatives, the subjects responded by pressing one of two telegraph keys mounted on a wooden board. Throughout the experiment, the left key was for "minor" responses and the right key was for "major" responses. Subjects responded with the index finger of the dominant hand, which was returned to rest on a stationary, central button after each response.

In the oddity discrimination task, in which there were three response alternatives, a different response board with three buttons was used. The subject's responses were registered and stored on disc file by an on-line computer system consisting of a GT-40 and PDP 11/45 in tandem.

Procedure

Music Questionnaire. Subjects filled out a questionnaire that asked them to specify in detail their current and past involvement with music and to rate themselves on various aspects of music ability on a scale from 1 to 10. Questionnaires were filled out before testing.

Initial Training. Subjects first listened to the prototype display tape. Next they were taught how to respond and then were asked to listen to the prototype display tape and respond with the appropriate key press. If the subjects expressed difficulty in identifying the stimuli, they were given additional training. Such subjects heard the display tape again, and the experimenter indicated after each chord whether it was minor or major. If further training was required, the subject was played about 15 chords from the prototype test tape (which were in random order), again with the experimenter verbally identifying each chord.

After this initial training, the subjects listened and responded to the complete prototype test tape again and identified each chord as minor or major.

Identification of the Continuum. The subjects were then told that they would hear more kinds of chords, and that the task would be to determine whether each chord was most like a minor or major chord. They were instructed to guess if they could make a decision on no other basis. The subjects then responded to the identification practice tape, followed by the identification test tape. A five minute rest period was allowed after the identification task had been completed.

Discrimination. Before the discrimination task, the subjects were told that they would hear triplets of chords, two of which were identical and one of which was different. They were instructed to press the button marked #1, #2, or #3, depending on whether the odd stimulus was first, second, or third

in the triplet. The subjects were told to guess if they could make a decision on no other basis. The subjects then listened to and responded to the discrimination practice tape followed by the discrimination test tape.

RESULTS AND DISCUSSION

Division into Subject Groups

The distribution of all subjects' ability to identify the prototype stimuli is displayed in Figure 2. All of the professional musicians could identify the prototype stimuli very accurately, ranging from 94 to 100 percent accuracy (mean = 97.6 percent). These subjects also rated themselves in the 8 to 10 point range on the 10-point scale of overall musical ability (mean rating = 9.4).

The results for the nonprofessional subjects plainly suggested a *post hoc* subdivision of this group into two groups, since the distribution of their ability to identify the prototype stimuli is clearly bimodal. Six of these subjects could identify the prototype stimuli quite accurately, ranging from 83 to 100 percent correct (mean = 90.3 percent). This group will be referred to as the NP-H group (nonprofessional, high-skill). This classification is based solely on their skill in identifying the stimulus prototypes. (When the professional musicians and the NP-H subjects are referred to together, the term "musically-skilled subjects" will be used.)

Five of the nonprofessional subjects had great difficulty identifying the prototype stimuli and performed just slightly above chance (mean = 54 percent). This group will be referred to as the NP-L group (nonprofessional, low-skill).

Although the classification of the NP-H and the NP-L groups was based solely on the ability to identify stimulus prototypes, this objective measure tended to match the subjects' self-reported estimates of musical ability and musical experience. The mean self-rating of the NP-H group on overall musical ability was 6.0, with all the subjects in this group rating themselves 5 or above, except one who rated herself 4. All of the subjects in this group had studied a musical instrument for 4 or 5 years, except one who had never studied an instrument. Of the five people in the NP-L group, all rated their musical ability as 3 or 4 (mean rating = 2.6). None had ever studied a musical instrument, except one student who had studied for five years.

In summary, then, the NP-H and NP-L groups each contained one subject whose experience with musical lessons suggested placement in the opposite group. However, it was decided that the ability to identify the prototype minor and major chords should be the only criterion used to separate the two groups, since this ability was the only direct measure of musical performance available. Nevertheless, this ability was highly correlated with various factors assessed by the Music Ability questionnaire: overall musical ability ($r = .79$, $p < .01$); number of years of music study ($r = .62$, $p < .02$); and number of hours spent playing an instrument each week ($r = .56$, $p < .05$). (See Table 1 for a complete listing of correlations, and Figure 3 for a

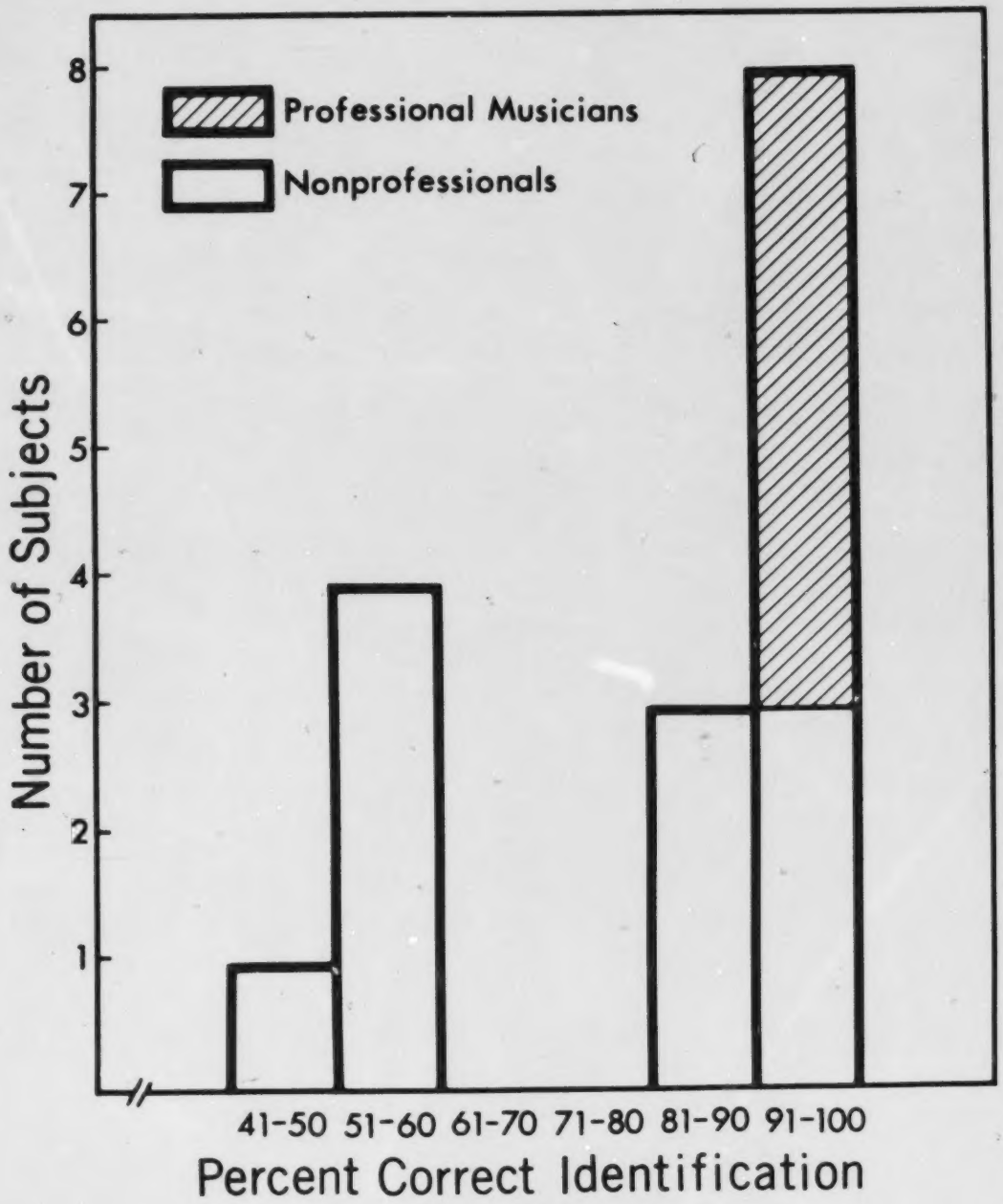

Figure 2: Distribution of scores for the identification of minor and major prototype chords.

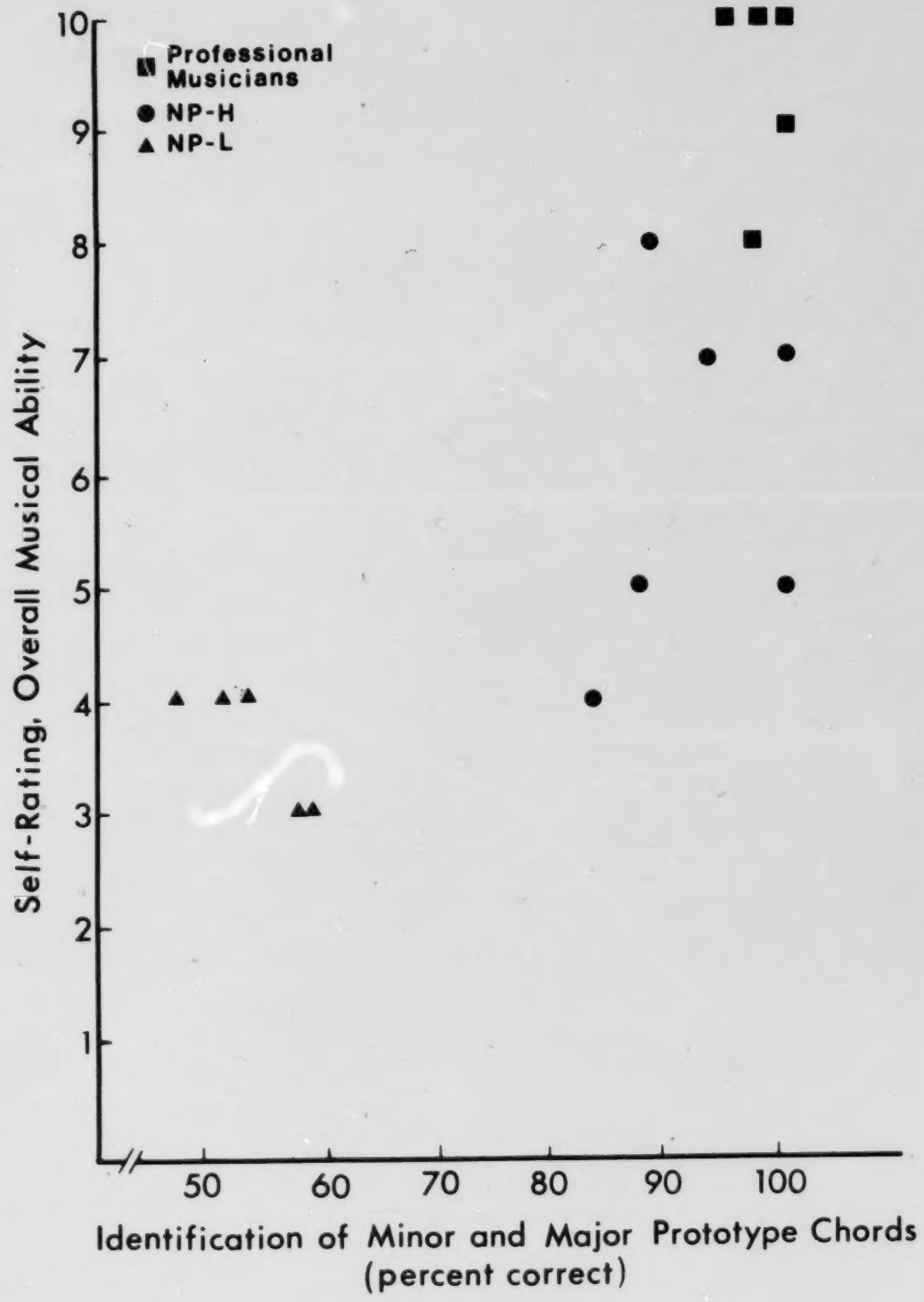

Figure 3: Relation between self-ratings of musical ability and identification accuracy of chord prototypes.

scatterplot of self-ratings of overall musical ability and accuracy in identifying the stimulus prototypes.) The factor of "hours spent listening to music" did not correlate with identification performance. "Listening," however, could have implied situations as diverse as leaving a radio on all day while doing something else, as well as listening carefully to a musical performance. It apparently did not necessarily imply attentive listening or understanding; one subject, whose identification accuracy was only 52 percent, reported listening to music 53 hours per week.

TABLE 1: Correlation between identification of chord prototypes and musical experience.

	Identification of Prototypes	Musical Ability	Years Studied Instrument	Play Music Hours per Week
Identification of Prototypes				
Musical Ability	.78**			
Years Studied Instrument	.62*	.84**		
Play Music Hours per Week	.56*	.83**	.96**	
Listen to Music Hours per Week	-.22	-.03	-.01	.12

**$p < .01$
*$p < .05$

Identification

Identification data for all the groups are shown in Figure 4 for the continuum of chord stimuli. The professional musicians and the NP-H group yield sharply demarcated identification functions that are considered a necessary (but not sufficient) characteristic of categorical perception, as discussed above. Throughout most of the stimulus continuum, the professional musicians show nearly 100 percent consistent identification of the stimuli. The NP-H subjects seem slightly less consistent, particularly at stimulus #1 and #2, but overall, their identification functions still appear quite

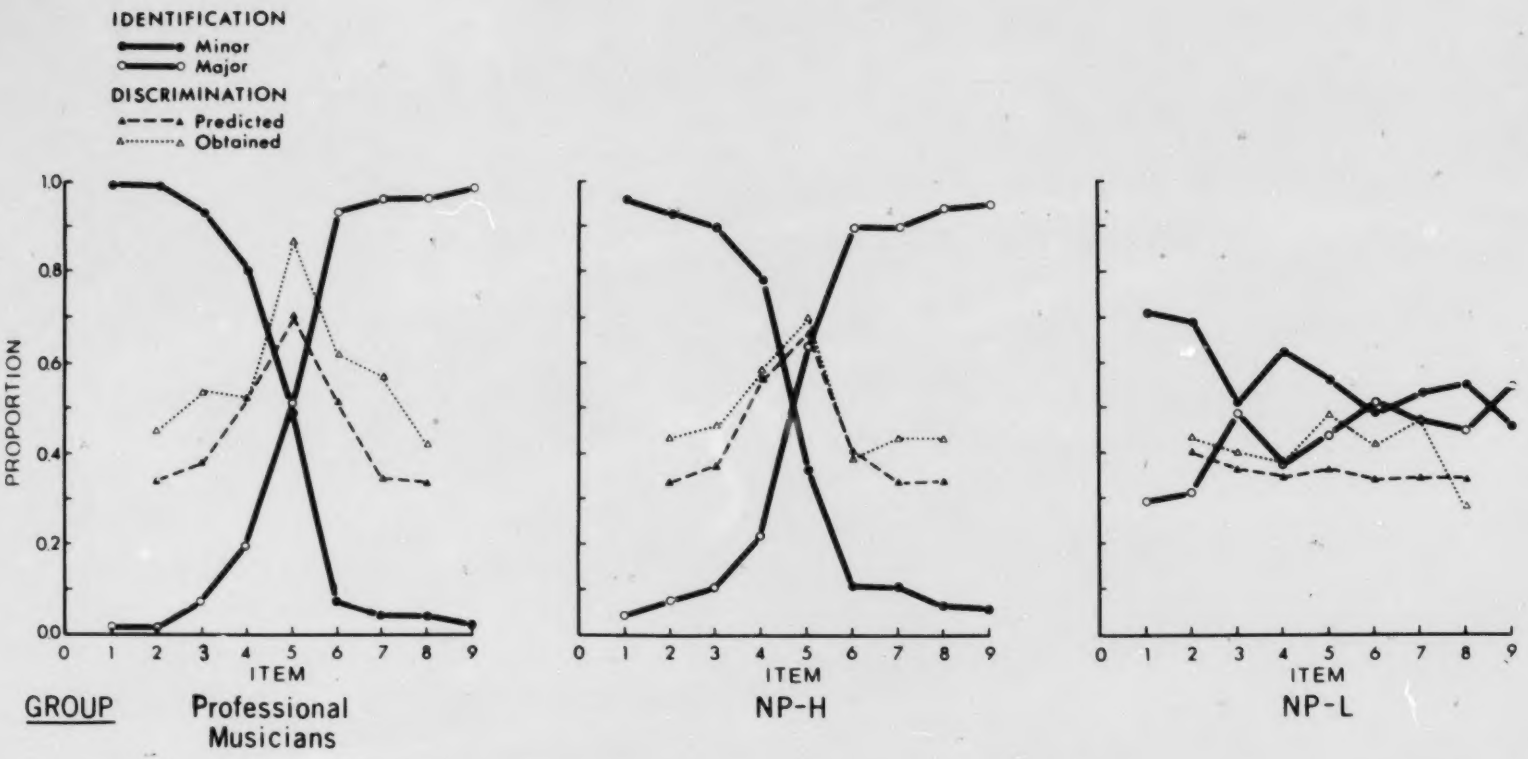

Figure 4: Identification and two-step discrimination of chords.

consistent. The crossover point from minor to major in both groups occurs at stimulus #5.

In the early experiments of the Haskins Laboratories group on categorical perception (for example, Liberman et al., 1957), subjects were preselected for their ability to identify speech stimuli consistently. This selection was done by visual examination of the data. A more precise, quantitative measure was substituted by Pisoni (1971). He estimated informational uncertainty for each stimulus along the continuum, and then converted these values so that complete response diversity (three equiprobable responses) would register as 0 and complete response consistency would equal 1.0.

A comparable measure of identification consistency has been applied to the present data (where only two response alternatives exist). The response consistency measures for each subject are displayed in Table 2. A Mann-Whitney test demonstrates that the higher consistency of the professional musicians relative to the NP-H subjects is marginally significant ($U = 6$, $p < .05$). Both of these musically skilled groups, however, are clearly more consistent than the NP-L group ($U = 0$, $p < .005$).

The present data for musical stimuli were compared with those for speech. Pisoni's (1971) subjects showed an average response consistency of .76 for stop consonants and .73 for steady state vowels. These values are comparable to the present data of the NP-H group. The professional musicians, however, appear to be more consistent in identifying chords than Pisoni's subjects were with either kind of speech stimulus. Whether this is due to the professional musicians' vast experience with musical stimuli, or to their possibly higher auditory acuity in general, cannot be determined from the present data.

Discrimination

Data for the discrimination of stimuli two steps apart on the stimulus continuum are displayed in Figure 4. The two musically-skilled groups again show the kind of functions that are characteristic of categorical perception. The professional musicians, in particular, show a high peak at the category boundary and troughs within the category. The NP-H subjects show a similar pattern, although their peak at the category boundary is less high (69 percent accuracy for the NP-H subjects vs. 87 percent for the professionals, $t(9) = 1.95$, $p < .05$). The NP-L subjects, on the other hand, show a noisy function that is clearly not categorical, with no consistent peak and performance only slightly above chance for most points along the continuum.

Affirmation of categorical perception requires a peak at the category boundary, but such a peak has not been defined statistically in most of the literature to date. For the present data, a statistical test of "peakedness" was devised by contrasting an across-boundary stimulus pair with the within-boundary pairs. There were seven possible pairings of items in the two-step discrimination task (1-3, 2-4, 3-5, 4-6, 5-7, 6-8, and 7-9). The stimulus pair 4-6 represented a definite across-boundary pair, while other pairs were clearly within-category discriminations (1-3 and 2-4 identified as minor, 6-8 and 7-9 as major). The pairs 3-5 and 5-7 were not included in this contrast

TABLE 2: Response consistency values in absolute identification.

Group	Subject	Chords	Single Tones
Professional Musicians·	RE	.86	.75
	JG	.83	.82
	DR	.74	.88
	SR	.79	.89
	EP	.96	.78
	Mean	.836	.824
NP-H	GG	.82	.67
	SG	.91	.83
	JD	.75	.69
	AM	.54	.88
	GC	.72	.79
	SS	.85	.69
	Mean	.765	.758
NP-L	SN	.26	.88
	WF	.29	.70
	HH	.33	.68
	GB	.36	.79
	CP	.18	.88
	Mean	.284	.786

because their status as within- or between-category pairs could vary depending on the individual subject's identification function. Thus, a minimum condition for statistically significant peakedness involved a contrast of the 4-6 pair against four within-category pairs, with an analysis of variance using the methods of weights described by Winer (1971).

The peakedness contrast, performed on the present data, shows a statistically significant peak for the professional musicians, $F(1,4) = 98.18$, $p < .001$, and for the NP-H subjects, $F(1,5) = 53.48$, $p < .001$. The same contrast performed on the NP-L data is not significant.

Expected and Predicted Functions

A further criterion for categorical perception is that the obtained discrimination functions match a predicted discrimination function that is derived from the identification data. If one assumes that the subjects are making the discriminations only on the basis of binary labels that they assign to the different stimuli, then for the oddity-discrimination task in this experiment, discrimination performance can be predicted by the following formula (Miyawaki et al., 1975):

$$P_{correct} = [\ 1 + 2(P_{min} - P_{min'})^2\ ] / 3$$

where P_{min} and $P_{min'}$ represent the proportions of labeling the two kinds of stimuli to be discriminated as minor.

For the present data, predicted discrimination functions were computed for each subject. These functions, averaged for each group, are compared with the obtained functions in Figure 4. The NP-H subjects show a very good fit between predicted and obtained functions. The professional musicians also show a good fit, although obtained performance seems to be somewhat higher than predicted at some points. (This difference is significant for two of the five subjects, as discussed below.) This occasional superiority of obtained over predicted performance is not uncommon in experiments with speech as well (for example, Liberman et al., 1957; Miyawaki et al., 1975). For the nonprofessional musicians, obtained and predicted functions are also not significantly different, but this result is not very surprising, since it essentially represents a close fit between predicted and obtained chance performance.

Goodness-of-fit between obtained and predicted functions was tested statistically for each subject, by calculating χ^2. Values of χ^2 for each subject are displayed in Table 3. For the professional musicians, three out of five show functions that do not differ significantly from the predicted functions. The two subjects whose functions do differ significantly show higher overall performance than predicted, but the shape of their obtained and predicted functions are highly correlated ($r = .79$, $p < .01$). For the NP-H group, five out of six of the subjects show a nonsignificant difference between obtained and predicted functions. All of the NP-L subjects yield nonsignificant values of χ^2 since they were expected to perform at chance on the discrimination task and did so.

TABLE 3: Goodness-of-fit of obtained and predicted two-step discrimination functions (χ^2)

Group	Subject	Chords	Single Tones
Professional Musicians	RE	23.321**	50.132**
	JG	9.798	68.496**
	DR	8.278	60.266**
	SR	22.307**	17.458*
	EP	13.520	32.262**
NP-H	GG	26.516**	46.926**
	SG	7.569	11.288
	JD	5.051	18.027*
	AM	5.496	9.585
	GC	1.468	7.264
	SS	10.542	16.655*
NP-L	SN	4.500	12.186
	WF	3.964	49.693**
	HH	11.208	6.589
	GB	7.125	12.595
	CP	12.000	13.455

**$p < .01$
*$p < .05$
All other comparisons, $p < .05$. For all comparisons, df = 7.

One-Step Discrimination of Chords

The results for the one-step discrimination trials are similar to those for the two-step discrimination, although less conclusive. As expected, the peaks in the discrimination functions are diminished. This is mainly because one-step discriminations cannot completely straddle the category boundary for this set of stimuli. The statistical measure of peakedness is significant for the professional musicians, but at a lower level of confidence than for the two-step function, $F(1,4) = 14.2$, $p < .05$. For the NP-H and NP-L subjects, the peakedness contrast is not significant for the one-step data.

The goodness-of-fit tests for the one-step task are generally also less conclusive. (See Table 4). Values of χ^2 are significant for only one NP-L subject ($p < .05$) and for two NP-H subjects ($p < .01$). Three of the professional musicians show significant values of χ^2 ($p < .05$), largely because of greater-than-predicted performance, as in the two-step discrimination trials.

Summary of the Chord Discrimination Data

Based on the identification and two-step discrimination functions, the NP-H subjects show categorical perception of the musical triads used in this experiment, in which the central tone varies from approximately 6/5 to 5/4 of the frequency of the lowest tone, that is, from a minor to a major third. The data of the professional musicians also appears categorical, with certain qualifications. Their sharp peak in discrimination at the category boundary shows that they can use the minor-major distinction to enhance discrimination. In fact, they appear to use labeling to make discriminations even more efficiently than the NP-H subjects. However, two of the professional musicians tend to discriminate two-step stimulus pairs slightly better than one would predict from labeling alone, which suggests that they can detect some purely auditory differences to enhance their discrimination capabilities. It cannot be determined from the present data, however, whether these subjects are especially sensitive to auditory differences only in musical stimuli with which they are unusually familiar, or whether they possess a heightened auditory acuity that would yield exceptionally high performance with speech as well.

Mandatory and Optional Processes in Discrimination Based on Analysis of Order Effects

One factor in the oddity discrimination task that has not been pursued in the literature is the effect of the temporal position of the "oddball" in the stimulus triad. Two important questions seem immediately apparent: a) Is discrimination performance equivalent for the three possible oddball positions? b) If not, does the shape of the discrimination curve interact with the oddball stimulus position?

The data relevant to this first question are displayed in Figure 5. The overall discrimination performance for the NP-L subjects, as noted above, does not differ significantly from chance. Yet there is a statistically reliable difference in performance between trials when the oddball stimulus is in the third position, as opposed to when it is in the first or second positions,

TABLE 4: Goodness-of-fit of obtained and predicted one-step discrimination functions (χ^2).

Group	Subject	Chords	Single Tones
Professional Musicians	RE	6.375	12.429
	JG	17.336*	7.500
	DR	17.336*	20.415**
	SR	10.875	9.868
	EP	16.875*	16.961*
NP-H	GG	23.544**	9.311
	SG	10.500	13.580
	JD	7.500	8.426
	AM	4.470	11.079
	GC	1.470	14.164
	SS	20.569**	13.500
NP-L	SN	13.092	7.842
	WF	11.496	6.996
	HH	12.343	6.996
	GB	16.746*	12.236
	CP	9.750	10.339

**$p < .01$.
*$p < .05$.
All other comparisons, $p < .050$. For all comparisons, df = 8.

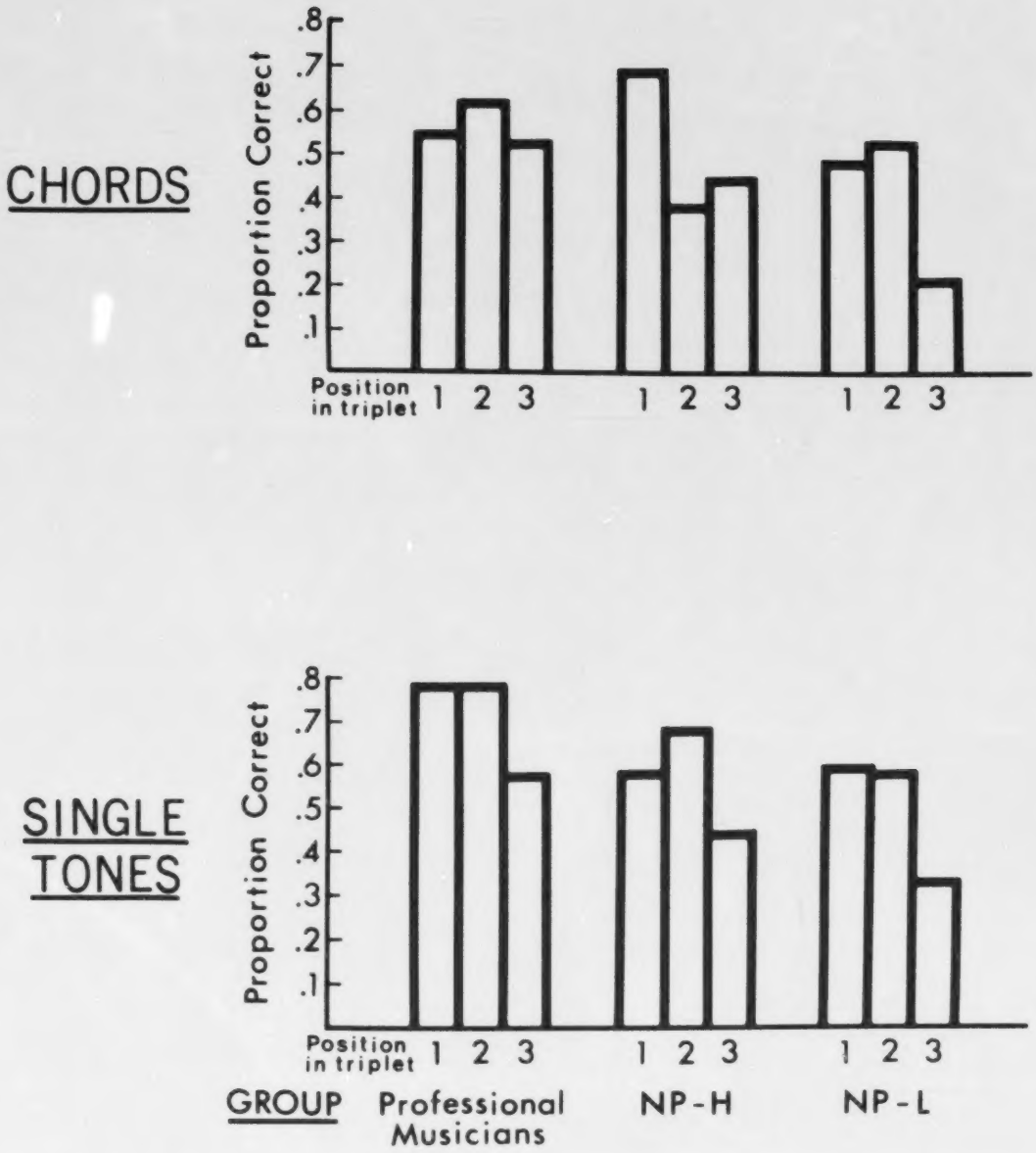

Figure 5: Effect of oddball position in discrimination triplet.

$F(2,8) = 16.63$, $p < .01$. (The difference between the first and second positions is not significant.) When the oddball is in the third position, discrimination performance is only 21 percent, which is below chance. This suggests a basic feature of the oddity discrimination task that has not been previously reported. Apparently, even when subjects are doing little more than guessing, there is a response bias away from the final position. One explanation for this pattern could involve attentional factors. Noncategorical comparisons of these stimuli based on their physical characteristics, that is, "analog" comparisons, take time. When the third of the three stimuli is presented, the subject is still involved in comparing the first two stimuli, and so does not attend adequately to the third stimulus. Since the subject is less likely to attend fully to the third stimulus, he is more likely to guess that the oddball is the first or second stimulus. This theory would predict that a similar order pattern will occur in any discrimination task in which the subject is making noncategorical, "analog" comparisons of the stimuli. (See the data of Experiment II for confirmation of this prediction.)

The performance of the professional musicians is about equal whether the oddball is in the first, second, or third position in the stimulus triad. In an analysis of variance, none of the small differences between positions is statistically significant. This fact is not surprising, since these subjects are probably relying almost completely on the categories of the stimuli for the discrimination task. Thus they are making relatively rapid, categorical, "digital" comparisons that allow them to attend adequately to the third stimulus.

For the NP-H subjects, surprisingly, performance is much higher for the first oddball position than for the final two. This observation is confirmed by an analysis of variance, which shows the effect of oddball position to be significant, $F(2,10) = 12.09$, $p < .05$, and by the *post hoc* comparisons which demonstrate that the two final positions are significantly different from the first position but not from each other.

To account for the order effect of the NP-H group, one must look even more closely at the data. Table 5 shows the accuracy data broken down into groups, stimulus comparison, and oddball position. Each of these numbers reflects the average of only four trials per subject, so one may expect the data to be somewhat unstable. Nevertheless, an extremely interesting pattern emerges. For the professional musicians, performance peaks at the category boundary, regardless of oddball position. Apparently, the strategy of these subjects is always to use the minor-major categories to an equivalent degree, regardless of oddball position. For the NP-H subjects, on the other hand, the peaks and troughs of the discrimination curve appear only when the oddball is in the second or third position. When it is in the first position, discrimination is above 60 percent at all points except for the 6-8 stimulus comparison. This pattern holds up under statistical scrutiny. A partitioning of the interaction term, contrasting the first with the second and third oddball positions, reveals that the Stimulus X Position interaction is highly significant, $F(1,5) = 28.52$, $p < .01$. The data for the professional musicians (as well as for the NP-L subjects) show no significant Stimulus X Position interaction.

TABLE 5: Discrimination accuracy (group X stimulus pair X oddball position).

Group	Oddball Stimulus Position	1-3	2-4	3-5	4-6	5-7	6-8	7-9
Professional musicians	1	.50	.65	.40	.85	.60	.40	.50
	2	.45	.50	.55	.90	.75	.75	.45
	3	.40	.45	.60	.85	.50	.55	.30
NP-H	1	.67	.75	.71	.75	.63	.38	.63
	2	.29	.29	.54	.58	.25	.38	.33
	3	.33	.33	.50	.75	.29	.54	.33
NP-L	1	.50	.55	.55	.45	.50	.55	.25
	2	.50	.55	.45	.60	.60	.60	.45
	3	.30	.10	.15	.40	.15	.25	.15

Figure 6 outlines the decision processes that may account for the position effect observed with the NP-H group. It is striking that the NP-H subjects can discriminate within-category auditory differences if the oddball is in the first position. But when the third stimulus of the discrimination triad is presented, and there is insufficient attention available for auditory comparisons, these subjects then resort to stimulus categories to perform the discrimination task.

It should be noted that the model also makes an auxiliary, quantitative prediction, that there will be a greater number of false alarms for the first oddball stimulus position than for the second or third positions. When the oddball stimulus is in the second position, the model predicts that the subject will perceive by auditory, "analog" comparison that the first and second stimuli are different. The second and third stimuli, however, will be compared only "digitally," that is, by categories, and will therefore seem to be identical. The subject will therefore respond that the stimulus is in the first position. This will result in a false alarm for the first stimulus.

The false alarm rate of the NP-H subjects was therefore examined. For the first, second, and third oddball positions, the average number of false alarms for the NP-H subjects are, respectively, 20.7, 12.5, and 10. Thus, the false alarm rate for the first position is approximately double that of the second or third positions. This finding was upheld statistically. An analysis of variance of these false alarm data reveals that the factor of oddball stimulus position is significant, $F(2,10) = 7.83$, $p < .025$. Subsequent comparisons by the Newman-Keuls method indicate that the first position is different from the second and third ($p < .05$), but that the second and third positions are not significantly different from one another.

It would seem then that the professional musicians, unlike the NP-H subjects, seem compelled to use category labels, even when such a strategy is

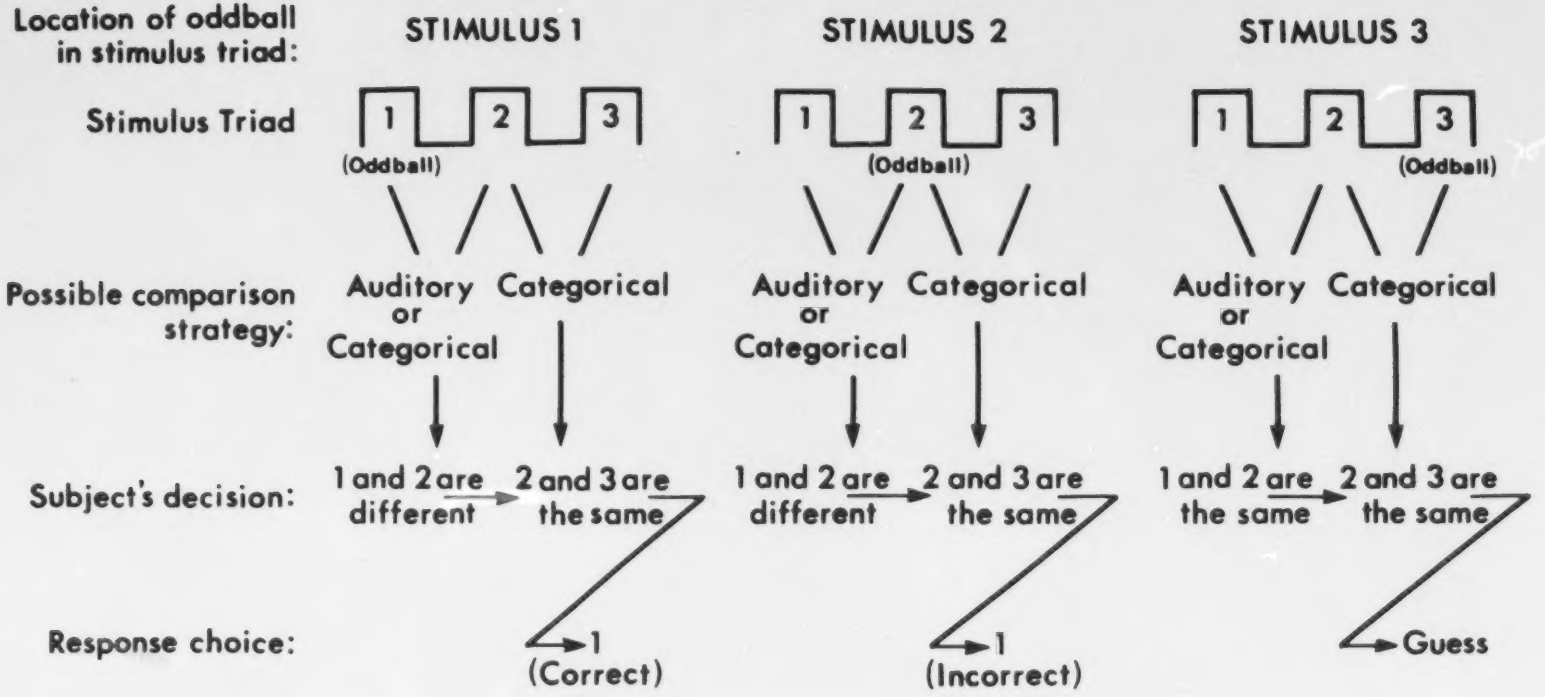

Figure 6: Decision strategy possibly used by NP-H subjects in the discrimination task when the stimulus comparison is within-category.

not optimal. Thus, this group does not perceive within-category auditory differences better when the oddball is in the first position. How can we account for this difference between the two groups? One way might be to think in terms of learning. For the professional musicians, the categorization of minor and major is overlearned; they use this distinction constantly in their work. For the NP-H subjects, the distinction between minor and major is known, but not used so much as to be overlearned.

If the present results were due solely to learning, then one might suspect that a different mechanism is involved in categorical perception of steady-state sounds, such as chords, and the categorical perception of sounds that vary rapidly over time, such as stop consonants or plucks and bows. The reason for this suspicion would be that stop consonants and plucks and bows are perceived in a categorical fashion by neonates (Eimas et al., 1971; Jusczyk et al., 1977), which, as noted above, suggests that innate neural tuning rather than learning is involved. Data concerning how neonates perceive the stimuli used in the present experiment would be especially useful in specifying the role of experience in chord perception.

Implications for the Interpretation of Previously Collected Data

The detailed analyses of the present data, examining discrimination curves for each of the three possible oddball positions, greatly clarifies the processes underlying the overall discrimination curves. However, they also lead one to be cautious in interpreting previous identification and discrimination experiments. One wonders, for example, whether there might have been any oddball position effects in the data of Mattingly, Liberman, Syrdal and Halwes (1971), who compared discrimination of consonant-vowel syllables and isolated second-formant transitions (chirps). One would expect the speech sounds to produce no significant order effect, and the chirps to show higher accuracy for the first and second position than for the third position (as in the noncategorical perception by the nonmusicians in the present experiment). One would expect no subjects to show the order pattern of the nonprofessional musicians, since, for native speakers, stop consonants are overlearned, and chirps are unfamiliar. Neither kind of stimulus, stops or chirps, is "partially learned."

It would also be interesting to perform the analysis of Stimulus-pair X Oddball-position on data from the studies of identification and discrimination of F_0 variation in tone languages. Klotz[1] found categorical perception for tonemes in native speakers of Mandarin Chinese, and noncategorical perception for English speakers who did not speak Chinese. However, he also tested a group of subjects who had studied Mandarin but whose native language was English, and they produced a peculiar discontinuous discrimination curve. Similarly problematic results were obtained by Chan, Chuang, and Wang (1975). The relationship between the Mandarin-speaking, native English subjects and the toneme stimuli may be similar to the relationship between the nonprofes-

---

[1]Klotz, R. (1975) A test for categorical perception of tone in Mandarin. Unpublished manuscript.

sional musicians and the chords in the present data. That is, the categories are known but not overlearned. Unfortunately, neither of the toneme studies reports having inspected position effects.

EXPERIMENT II

Although both musical groups (professional and NP-H) appear to perceive the chords categorically in Experiment I, the exact mechanism underlying this result is unclear. Three alternatives seem possible: a) there is a special "harmonic mode" of perception, similar to the speech mode, which ignores acoustic differences not relevant to the harmonic distinctions possible in music based on the tempered scale; b) although the pitch of the central tone is varied stepwise in a continuous fashion, there is an intrinsic perceptible discontinuity in the stimulus array, based on some acoustic interaction between the component tones of the chord; c) the discontinuity in chord perception is due to a discontinuity in simple pitch discrimination. Although this last proposition seems unlikely, (given other research), it must be treated empirically before considering the other two explanations of categorical perception with chords. Therefore, the same subjects of Experiment I participated in another set of identification and discrimination tasks, but the stimuli were the isolated central tones of the chords, that is, the "minimal cues" to the major-minor distinction.

METHOD

Stimuli

The nine stimuli used in this experiment were sine waves 250 msec in duration, ranging in 2.32 Hz steps from 311.1 to 329.6 Hz. These stimuli were identical to the central tones of the chord stimuli in Experiment I.

Subjects, Tapes and Apparatus

The 16 subjects in this experiment were the same as those who had participated in Experiment I about a week earlier. All subjects participated in Experiment II after Experiment I; although counterbalancing would have been desirable, it was not feasible because of the small number of available professional musicians who met the criteria given in Experiment I.

The tapes used in this experiment were identical in design to those of Experiment I. The only difference was that the tapes contained only the central tone of each chord instead of the complete three-tone chords. The apparatus was identical to that of Experiment I.

Procedure

The procedure was virtually identical with Experiment I, except that there was no variation in the training procedure, since all subjects could readily identify the single prototype tones (extracted from chords #1 and #9) as "low" or "high." In the identification task of this experiment, the left telegraph key was always used for the "low" responses, and the right key was always used for "high" responses.

Ideally, in Experiments I and II, there would have been complete counterbalancing of the order of presentation of the two kinds of stimuli (chords and single tones), and of the order of tasks (identification and discrimination) within each kind of stimulus. However, because of the difficulty in recruiting a homogeneous sample of professional musicians, this ideal could not be achieved. Although it is unlikely that substantially different results would have occurred with such counterbalancing, this limitation of the experimental design should be noted.

RESULTS AND DISCUSSION

Identification

All subjects are very accurate in identifying the extreme tones of the continuum (that had been part of the prototype major and minor chords). Mean accuracy in this task is 98 percent for each of the three groups.

Data from the identification of the entire stimulus continuum of single tones are displayed in Figure 7. The functions of all groups, including the NP-L subjects, are quite well demarcated. However, the crossover point is more variable than was the case with chords, occurring at stimulus #5 for NP-L and NP-H subjects, and at stimulus #4 for the professional musicians.

Response consistency values are displayed in Table 2. As with the chords, consistency values for the professional musicians are slightly higher than for the NP-H subjects, although in this case the difference is not statistically reliable ($U = 8$, $p = .12$). Surprisingly, however, the NP-L subjects are also quite consistent in identifying single tones--in fact, slightly (but not significantly) more consistent than the NP-H subjects. This fact is important. The NP-L subjects clearly can perceive differences between individual pitches; they are not "tone deaf." Their poor identification of chords in Experiment I is more likely due to an inability to identify harmonic relations. Certainly, the context of two other fixed tones renders the perception of changes in a third tone more difficult. Both musically-skilled groups, however, can then rely on the overall harmonic gestalt that results from the tone combination. The NP-L subjects cannot. For them the outer two tones of the chord may represent only interference that masks the change in the frequency of the central tone.

Discrimination

Although all of the groups produce identification functions for single tones that are quite well demarcated, none of the discrimination functions for single tones fulfills all of the requirements for categorical perception. (See Figure 7 for a display of the two-step discrimination curves and Tables 3 and 4 for the values of χ^2 indicating the goodness-of-fit between obtained and predicted functions.) For all groups, the test of peakedness is not significant for either one- or two-step discrimination functions, with one exception: the NP-H subjects' curve is significantly peaked, $F(1,5) = 8.03$, $p < .05$. However, their performance tends to be higher than predicted, although the data across individual subjects are not entirely consistent: χ^2 values for one subject are significant, $p < .01$; for two subjects, $p < .05$; for the other

FIGURE 7

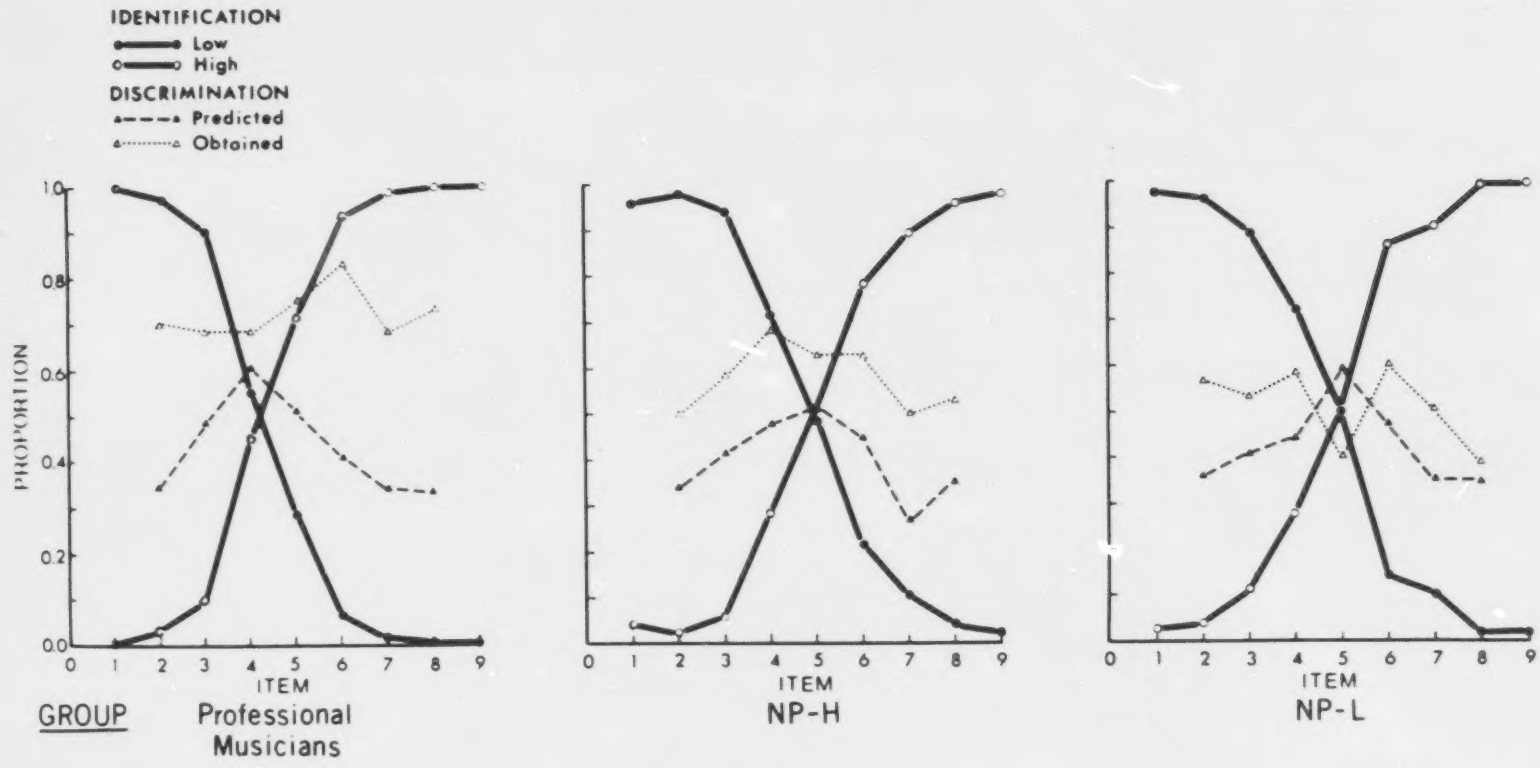

Figure 7: Identification and two-step discrimination of single tones.

three subjects, χ^2 was not significant. Moreover, Figure 7 shows that the maximum peak occurs not at stimulus #5 (the category boundary) but at stimulus #4. Thus the perception of single tones, even in the case of the NP-H subjects, does not satisfy two of the conditions for categorical perception mentioned above (nonsignificant goodness-of-fit between obtained and predicted functions, and the coinciding of the discrimination peak with the labeling boundary).

The parallels between these results and those for speech perception are quite compelling. In the present work, the single tones, which form the "minimal cues" for the minor-major distinction, are not in themselves perceived categorically. Thus, categorical perception of chords seems to involve processing the gestalt produced by the tones in the chords. Similarly, categorical perception of consonant-vowel syllables is produced by the gestalt of the entire formant structure, not by the isolated second-formant transitions that are the "minimal cues" for phonetic distinctions between stop consonants (see the data of Mattingly et al., 1971).

Order Effects in the Discrimination of Single Tones

The effect of oddball position in the discrimination of single tones is displayed in Figure 5. All three groups show the same pattern: equivalent performance for position 1 and 2 and sharply decreased performance for position 3. Statistical analyses by Newman-Keuls comparisons confirm the reliability of this pattern ($p < .05$ for all three groups). This order effect may characterize oddball discrimination of any stimulus array, whenever discrimination is not based on labeling categories. When discrimination is based entirely on labeling categories, however, there is equivalent performance in all three oddball positions. It is apparent, in summary, that an analysis of order effects can substantially clarify the processes underlying discrimination curves that are derived from collapsing over all trials for a stimulus pair.

GENERAL DISCUSSION

The Inability to Perceive Harmonic Distinctions

Several results of significance have emerged from the present study. The first is that approximately a third of the subjects surprisingly could not discriminate between the prototype minor and major chords used in this study. One might have expected that people without musical training would not know how to label chords as minor or major. However, the NP-L subjects had a more comprehensive deficit: they could not perceive any difference between the two types of chords, even without any labeling requirement.

The inability of such a sizable proportion of subjects to perceive differences in a dimension so basic to the organization of tonal music as minor-major may have serious consequences for the way certain people listen to music. However, two reservations must be stressed concerning the generalizability of the present results. First, the subject sample used in the present study was not randomly collected. The five subjects who could not identify the stimulus prototypes had responded to a sign-up sheet asking for "nonmusi-

cians," so that subjects lacking in musical ability may consequently have been overrepresented in the sample. It is possible, therefore, that the true proportion of the population at large who cannot perceive the minor-major distinction in this task is smaller.

A second reservation about the generalizability of the present results is that the stimuli used in this experiment differed in several significant ways from naturally-produced musical sounds. They were steady-state sine waves with no overtones. The lack of overtones may diminish the perceived difference between minor and major chords, since the higher partials may interact differently in the two kinds of chords. For example, the fourth partial of the note "C" is approximately E-natural. In a C-major chord, this partial is nearly coincident with the third partial of the third (E), but in a C-minor chord the fourth partial of C is discordant (that is, produces beats) with the third partial of the third (E-flat). Therefore, the subjects who could not discriminate the synthetic minor and major chords might perform more accurately with naturally-produced musical stimuli that possess an overtone structure.

Clearly, a study is needed to determine more accurately what proportion of the population cannot perceive the difference between minor and major chords in naturally-produced music. Should further research support the surprising finding of the present study, it would suggest several important questions. If certain people do not perceive basic harmonic distinctions, what do they perceive in music? (Remember that one of the NP-L subjects "listened" to music 53 hours per week.) A person who cannot perceive the difference between minor and major probably would also not appreciate the significance of augmented sixth and diminished seventh chords, or other more complex harmonic relationships that form much of the substance of serious music. If the inability of much of the population to distinguish minor from major were established as a fact, it might suggest a reason why popular music is often harmonically very simple and depends in so large part on rhythmic devices.

Categorical Perception of Steady-State Stimuli

The mere fact that categorical perception has been found to occur with the chords used in the present study is of interest, especially when one considers the result in relation to the various current theories of the mechanisms underlying categorical perception. As noted above, it was originally thought that categorical perception was a phenomenon unique to speech stimuli, but this hypothesis has repeatedly not been supported in recent years by findings of categorical perception with nonlinguistic sounds. Alternative explanations have since been proposed that stress a neural limitation in the ability to resolve temporal variation as a basis for categorical perception. Such theories seem to account for the categorical perception of various auditory and visual stimuli whose distinctions are cued by rapid physical events. However, the neural limitation hypothesis, as currently formulated, does not account for the results of the present experiments, because the chords used in these experiments did not vary over time. They were steady-state, with respect to both pitch and intensity. A neural tuning hypothesis could conceivably be extended to the resolution of steady-state pitch relationships, but such a theory would require a good deal more knowledge and data

on the neurophysiological mechanisms that underlie the perception of harmonic relationships, comparable to the neurophysiological data on the perception of rapid temporal variation in pitch and intensity (for example, Møller, 1971). At present, it would seem more prudent to account for the categorical perception of chords in terms of learning concepts, such as "acquired distinctiveness" and "acquired similarity," that can apply to experience with music as well as with speech.

The Role of Experience in the Perception of Speech and Music

The findings of Experiments I and II with musical sounds present a striking parallel to the findings of Miyawaki et al. (1975) using speech sounds. (See Table 6 for a summary and comparison of their results and those of the present experiment.) Miyawaki et al., as noted above, found that speakers of English perceive the distinction between /r/ and /l/ categorically, while speakers of Japanese, a language in which the /r/-/l/ distinction is not phonemic, do not perceive the distinction categorically. In fact, the Japanese speakers for the most part simply cannot perceive the difference between /r/ and /l/. Yet, when the minimal acoustic cues of the /r/-/l/ distinction, the third formant (F_3) transitions, are presented in isolation, both groups produce equivalent results. Miyawaki et al. therefore conclude that linguistic experience is necessary for perception of linguistic units, such as phonemes, but that lack of linguistic experience does not hamper the perception of nonspeech sounds that comprise the minimal acoustic cues for the phonetic distinction.

The present results do not contradict the data of Miyawaki et al., but they do suggest a different generalized conclusion. Professional musicians and the NP-H subjects, experienced with the "language" of harmony, show categorical perception of the minor-major distinction, while the NP-L subjects (for the most part inexperienced with music) do not perceive the chord stimuli categorically. In fact, analogous to the Japanese speakers, the NP-L subjects simply do not perceive the distinction between minor and major chords. Thus, the present data suggest that the phenomenon observed by Miyawaki et al. may extend beyond language to perception of other encoded systems of sound. The role of experience in the perception of phonetic distinctions may be a special case of the role of experience in the perception of any acoustically complex stimulus that can be encoded within an organized system of sound, whether that system is linguistic or is, like music, nonlinguistic.

One problem in relating the present data to those of Miyawaki et al. deserves clarification. It may be objected that in the present experiment, the professional musicians show categorical results, but that performance of two of the subjects is slightly but significantly higher than predicted, suggesting an additional auditory component in their discrimination performance. Miyawaki et al., in discussing the results of their English-speaking subjects, note similarly that obtained performance is also slightly higher than predicted. Unfortunately, they do not report any statistical tests comparing obtained and predicted discrimination performance. Their averaged data from the English speakers seem to resemble the data of the professional musicians in the present experiment, but no precise quantitative comparison of the two sets of data is possible without individual χ^2 values.

TABLE 6: Summary and comparison of the present data with those of Miyawaki et al. (1975).

	Present experiments (music)			Miyawaki et al. (Speech)	
	Chords	Single tones		Syllables /ra/-/la/	F_3 transitions
Professional musicians and NP-H groups	Categorical perception	Noncategorical perception	English speakers	Categorical perception	Noncategorical perception
NP-L	Cannot perceive differences	Noncategorical perception	Japanese speakers	Cannot perceive differences	Noncategorical perception

More Than One Basis for Categorical Perception?

Although the learning explanation seems to account most readily for the present data on the categorical perception of chords, it does not rule out the involvement of innate neural mechanisms as well. In general, one can argue that the neural tuning and learning hypotheses are never mutually exclusive. In the area of categorical perception of speech sounds, in fact, there are data showing that innate mechanisms can be modified by experience. For example, the voice-onset time (VOT) boundary for native English speakers is approximately +25 msec (Lisker and Abramson, 1970), while for monolingual adult speakers of Spanish, the VOT boundary is -4 msec (Williams, 1974). Yet for Guatemalan infants raised in Spanish-speaking environments, four- to six-and-a-half months of age, there are two VOT boundaries, one between -20 and -60 msec, and the other between +20 and +60 msec, the latter corresponding to the boundary of native English speakers (Lasky, Syrdal-Lasky and Klein, 1975). Thus it appears that the VOT boundary common to English may be innate in infants of all nationalities, yet linguistic experience with certain languages such as Spanish can markedly alter this boundary. (See also Streeter, 1976, for similar findings with Kikuyu infants.)

It is also possible that different mechanisms may underlie the apparent categorical perception of different stimuli depending on their acoustic structure (that is, rapidly-varying vs. steady-state), or that different subjects may perceive the same stimuli in different ways, but these differences may be masked by averaged data. In the present study, the averaged data of NP-H subjects seem to satisfy the criteria of categorical perception better than the data of the professional musicians. Yet after examining the data more closely, we might come to a different conclusion. We find that in at least one oddball position, the NP-H subjects' discrimination data appears to be free of the effects of stimulus categories. The professional musicians, on the contrary, appear to use stimulus categories equally, regardless of oddball position. In an earlier day, only the overall functions might have been examined, and the differences between groups might have been interpreted as a difference in degree of categorical perception. The detailed analysis of order effects, however, has demonstrated that the difference between these groups may reflect a qualitative difference in the pattern of data. This is perhaps one of the most interesting findings of the present study, and it emphasizes the need for detailed analysis of all data from categorical perception experiments rather than the examination of only averaged data.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In Proceedings of the 5th International Congress of Acoustics, ed. by D. E. Commins. (Liege: Imp. G. Thone).

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. In Proceedings of the Sixth International Congress of Phonetic Sciences. (Prague: Academia), 569-573.

Chan, S. W., C. K. Chuang and W. S-Y. Wang. (1975) Cross-language study of categorical perception for lexical tone. Paper presented at the 90th Meeting of the Acoustical Society of America.

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner and C. F. Foard. (1976) Perceptual categories for musiclike sounds: implication for theories of speech perception. Quart. J. Exp. Psychol. 28, 361-378.

Eimas, P. D. (1975) Auditory and phonetic coding of the cues for speech: Discrimination of the (R-L) distinction by young infants. Percept. Psychophys. 18, 341-347.

Eimas, P. D., E. R. Siqueland, P. W. Jusczyk and J. M. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Fry, D. B., A. S. Abramson, P. D. Eimas and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Sp. 5, 171-189.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute 28. (Tokyo: University of Tokyo), 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute 29. (Tokyo: University of Tokyo), 207-214.

Hirsh, I. J. (1959) Auditory perception of temporal order. J. Acoust. Soc. Am. 31, 759-767.

Jones, M. R. (1976) Time, our lost dimension: Toward a new theory of perception, attention, and memory. Psychol. Rev. 83, 323-355.

Jusczyk, P. W., B. S. Rosner, J. E. Cutting, C. F. Foard and L. B. Smith. (1977) Categorical perception of nonspeech sounds by 2-month-old infants. Percept. Psychophys. 21, 50-54.

Kuhl, P. K. and J. D. Miller. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science 190, 69-72.

Lasky, R. E., A. Syrdal-Lasky and R. E. Klein. (1976) VOT discrimination by four and six and a half month old infants from Spanish environments. J. Exp. Child Psychol. 20, 215-225.

Liberman, A. M., K. S. Harris, P. Eimas, L. Lisker and J. Bastian. (1961) An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. Lang. Sp. 4, 175-195.

Liberman, A. M., K. S. Harris, H. S. Hoffman and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Liberman, A. M., K. S. Harris, J. A. Kinney and H. Lane. (1961) The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61, 379-388.

Lisker, L. and A. S. Abramson. (1970) The voicing dimensions: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences. (Prague: Academia), 563-567.

Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex 9, 355-369.

Mattingly, I. G., A. M. Liberman, A. Syrdal and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63, 81-

97.

Miller, J. D., C. C. Wier, R. E. Pastore, W. J. Kelly and R. J. Dooling. (1976) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Am. 60, 410-417.

Miyawaki, K. W., W. Strange, R. Verbrugge, A. M. Liberman, J. J. Jenkins and O. Fujimura. (1975) An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. Percept. Psychophys. 18, 389-397.

Møller, A. R. (1972) Coding of amplitude and frequency modulated sounds in the cochlear nucleus of the rat. Acta Physiologica Scandinavica 86, 223-238.

Pastore, R. E. (1976) Categorical perception: A critical re-evaluation. In Hearing and Davis: Essays Honoring Hallowell Davis, ed. by S. K. Hirsh, D. H. Eldredge, I. J. Hirsh and S. R. Silverman. (St. Louis: Washington University Press).

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished doctoral dissertation. University of Michigan.

Pisoni, D. B. (in press) Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing and stops. J. Acoust. Soc. Am.

Stevens, K. N., A. M. Liberman, M. Studdert-Kennedy and S. E. G. Öhman. (1969) Cross-language study of vowel perception. Lang. Sp. 12, 1-23.

Streeter, L. A. (1976) Language perception of 2-month old infants shows effects of both innate mechanisms and experience. Nature 259, 39-41.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Williams, L. (1974) Speech perception and production as a function of exposure to a second language. Unpublished doctoral dissertation, Harvard University.

Winer, B. J. (1971) Statistical Principles in Experimental Design, 2nd Edition. (New York: McGraw-Hill).

Phonetic and Auditory Aspects of Adaptation: Evidence from Thai Voicing Contrasts*

S. Lea Donald†

ABSTRACT

A series of six adaptation conditions using Thai subjects was carried out in order to investigate a possible phonetic contribution to adaptation effects on a voice onset time (VOT) continuum. Thai subjects divide a VOT continuum into three phonological categories at the labial and dental places of articulation, but only two phonological categories (voiceless inaspirates and voiceless aspirates) at the velar place. The effects of -70 msec labial and velar adaptors on both a labial and a velar VOT continuum were determined. Also, the effect of a 5 msec labial adaptor on the labial VOT continuum and the effect of a 5 msec velar adaptor on the velar VOT continuum were tested. The results of these experiments indicate the presence of both phonetic and auditory contributions to the adaptation effect.

INTRODUCTION

One aim of recent adaptation experiments has been to determine whether the boundary shifts obtained are due to alterations of normal auditory or phonetic processing. The experiments reported in Donald (1976) indicate that the phonological structure of a language may limit the effects of adaptation. English-speaking and Thai-speaking subjects were tested using labial stop stimuli varying along a voice onset time (VOT) continuum. Thai speakers divide a VOT continuum into three phonological categories (voiced, voiceless unaspirated, and voiceless aspirated stops). English speakers divide the same continuum into two phonological categories (voiced and voiceless stops). The series of labial stops used in this experiment thus spanned three phonological categories for the Thai-speaking subjects, but only two categories for the English-speaking subjects. Of greatest interest are the effects of the -80 msec and 5 msec adaptors. In Thai, where the two adaptors belong to different phonological categories, only the 5 msec adaptor significantly shifted the

---

*This is an expanded version of a paper presented at the 93rd meeting of the Acoustical Society of America, University Park, Pennsylvania, 6-10 June, 1977.

†Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

boundary between voiceless aspirates and voiceless inaspirates. In English, on the other hand, where the two adaptors belong to the same phonological category, both adaptors significantly affected the counterpart voiced-voiceless boundary. In short, the same acoustic signal, the -80 msec adaptor, affected subjects of different linguistic backgrounds differently. The main findings of this experiment were replicated by Foreit (1977). The data demonstrate that phonetic categorization can limit the observed effects of adaptation.

Furthermore, these findings can be interpreted as indicating that adaptation affects a phonetic level of processing, as discussed below. Nonetheless, the actual locus of adaptation may still be auditory rather than phonetic, since the observed effects may be due to differences in the way the two groups interpret the output of acoustic detectors. To illustrate, English speakers would compare the output of detectors sensitive to acoustic cues of voicing lag not only with the output of detectors sensitive to cues of coincident voice onset, but also with the output of detectors sensitive to cues of voicing lead in order to determine whether a stimulus was voiced or voiceless. Thus, a -70 msec labial stimulus would excite voicing lead detectors and the output of these detectors would be summed with the output of coincident voice onset detectors before being compared with the output of voicing lag detectors. Since the latter detectors would not have been excited by the -70 msec stimulus, the stimulus would be judged voiced rather than voiceless. Similarly, a 5 msec labial stimulus would excite coincident voice onset detectors and the output of these detectors would again be summed with the output of voicing lead detectors before being compared with the output of voicing lag detectors. Since the latter detectors would again not have been excited, the 5 msec labial stimulus, like the -70 msec labial stimulus, would be judged voiced rather than voiceless.

For Thai subjects, on the other hand, the -70 msec labial stimulus, excited by the output of voicing lead detectors, would be compared (rather than summed, as in the English case) with the output of coincident voice onset detectors in order to determine whether the stimulus was a prevoiced or a voiceless unaspirated stop. For the 5 msec labial stimulus, the output of the coincident voice onset detectors would be compared with both the voicing lead detectors and the voicing lag detectors. Since the output of coincident voice onset detectors would be stronger than the output of either of the other sets of detectors, the stimulus would be judged a voiceless unaspirated stop.

By this reasoning, adaptation of voicing lead detectors by repeated presentation of a -70 msec VOT stimulus would affect Thai and English speakers differently. For speakers of English, the voiced-voiceless judgment would be affected; for speakers of Thai, the prevoiced-voiceless unaspirated judgment would be affected. These are exactly the findings of Donald (1976) and Foreit (1977).

In order to test these alternative explanations of adaptation on the perception of voicing contrasts, it would be useful to find a parallel to the Thai-English condition, but within a single language, so that the same set of subjects can participate in both conditions. Thai, in fact, presents just

such a condition: although it uses three voicing categories among labial and dental stops, it uses only two voicing categories among velar stops: voiceless inaspirates and voiceless aspirates.

Table 1 outlines the boundary shifts predicted if adaptation depends on the phonological structure of a language. Thai subjects categorize a velar stop with voicing lead as a voiceless unaspirated stop; consequently, adaptation with a -70 msec velar stimulus should produce a significant boundary shift between voiceless inaspirates and voiceless aspirates on a velar series. By the same rationale, testing the labial series after presentation of this adaptor should also produce a significant boundary shift between voiceless unaspirated and voiceless aspirated stops. Unlike a velar stop with voicing lead, a labial stop with voicing lead is categorized by Thai speakers as a prevoiced stop. Consequently, testing either a velar or a labial series after adaptation with a -70 msec labial stop should produce no significant shift on the boundary between voiceless inaspirates and voiceless aspirates on either series.

TABLE 1: Boundary shifts between voiceless inaspirates and voiceless aspirates predicted by the phonetic hypothesis.

	Velar Series	Labial Series
-70 msec velar adaptor	shift	shift
-70 msec labial adaptor	no shift	no shift

Table 2 outlines the boundary shifts predicted if adaptation takes place solely at an auditory level of processing. By this hypothesis, when Thai subjects are asked to identify velar stops, they compare the output of detectors sensitive to cues of voicing lag not only with cues of coincident voice onset, but also with cues of voicing lead. Consequently, adaptation with either a velar or a labial stimulus with voicing lead should shift the boundary between voiceless inaspirates and voiceless aspirates on a velar series. In the perception of labial voiceless unaspirated stops, on the other hand, a comparison is drawn only between the output of detectors sensitive to cues of voicing lag and cues of coincident voice onset. Thus, neither a velar nor a labial stimulus with voicing lead should shift the boundary between voiceless inaspirates and voiceless aspirates on a labial series.

TABLE 2: Boundary shifts between voiceless inaspirates and voiceless aspirates predicted by the auditory hypothesis.

	Velar Series	Labial Series
-70 msec velar adaptor	shift	no shift
-70 msec labial adaptor	shift	no shift

Note that both hypotheses predict that the velar adaptor will produce a boundary shift on the velar series, while the labial adaptor will not produce a boundary shift on the labial series. The hypotheses yield opposite predictions on the cross-series conditions. The phonetic hypothesis predicts that the velar adaptor will shift the labial boundary, but that the labial adaptor will not shift the velar boundary. Conversely, the auditory hypothesis predicts that the velar adaptor will not shift the labial boundary, but that the labial adaptor will shift the velar boundary.

The within-series effects of velar and labial adaptors with coincident voice onset may also distinguish between the auditory and phonetic hypotheses. Both hypotheses predict that significant boundary shifts will occur in these two conditions. The voiceless unaspirated labial adaptor should shift the boundary between voiceless inaspirates and voiceless aspirates on a labial series. Donald (1976) and Foreit (1977) have verified this prediction. Likewise, a velar voiceless unaspirated adaptor ought to shift the boundary between voiceless inaspirates and voiceless aspirates on a velar series.

However, the two hypotheses differ in their predictions of the magnitude of the various boundary shifts. Various studies (for example, Bailey, 1973; McNabb, 1975; Miller, 1976) show that adaptation with clearly unambiguous exemplars of a phonetic class produces a larger effect than does adaptation with less acceptable stimuli. If adaptation is taking place only at a phonetic level of processing, the boundary shifts produced on a velar series by adapting with a voiceless unaspirated velar, a prevoiced velar, or a voiceless unaspirated labial should all be equal. If, on the other hand, adaptation is occurring at an auditory level of processing, the boundary shift produced by the velar voiceless unaspirated adaptor should be greater than the boundary shift produced by the velar prevoiced adaptor, since the prevoiced adaptor is not a good acoustic exemplar of a velar voiceless inaspirate. Likewise, the velar voiceless unaspirated adaptor should produce larger boundary shifts than the prevoiced labial adaptor, since there is more acoustic distance between adaptor and test series in the second instance. Table 3 summarizes these predictions.

TABLE 3: Relative magnitude of boundary shifts between voiceless inaspirates and voiceless aspirates on velar series predicted by auditory hypothesis.

-70 msec velar adaptor	<	5 msec velar adaptor	>	-70 msec labial adaptor

For the labials, the phonological hypothesis again predicts no difference in magnitude between the boundary shifts obtained from the prevoiced velar adaptor or from the voiceless unaspirated labial adaptor. An auditory account predicts that only the voiceless unaspirated labial adaptor will produce a boundary shift.

Subjects

Eight native speakers of Central Thai participated in all conditions. Three were students at Yale University and five were students at the University of Massachusetts at Amherst. All reported normal hearing and were naive as to the purpose of the experiment. Subjects were paid for their participation.

Stimuli

The stimuli used were synthetic labial and velar VOT series from the continua prepared by Lisker and Abramson (1970). The variations in VOT were produced by varying the onset of the first formant relative to the onset of the second and third formants. During the absence of the first formant, the upper formants were excited by a noise source rather than a periodic source. The VOT values range from -70 msec to 0 msec in 10 msec steps; from 0 msec to 50 msec in 5 msec steps; and from 50 msec to 70 msec in 10 msec steps. The adaptors used were -70 msec and 5 msec stimuli from both series. These values were chosen with the expectation that the -70 msec velar adaptor and both 5 msec adaptors would be identified as voiceless inaspirates, whereas the -70 msec labial adaptor would be identified as a prevoiced stop.

Procedure

Each subject participated in six experimental sessions. Each session consisted of an identification test followed by an adaptation test. In the first session, the identification test was 200 trials long, with each of the velar stimuli presented in random order ten times. The second session presented labial stimuli in an identification test that was also 200 trials long. In each of the subsequent sessions, a 100-trial identification test was presented, using labial and velar stimuli in alternate sessions. Altogether, each stimulus from both continua was identified 20 times by each subject. Within each test, the stimuli were presented in blocks of 15, with 3 seconds separating the stimuli within a block. Ten seconds separated the blocks.

Before each identification test, subjects were presented with alphabetic symbols (in Thai orthography) corresponding to the consonants they were to hear. Three of the Thai-speaking subjects responded with a number corresponding to their choice. The remaining Thai-speaking subjects responded with the orthographic symbol identifying the stimulus.

An adaptation test followed the identification test in each session. Subjects were exposed to 60 repetitions of the adapting stimulus with an inter stimulus interval (ISI) of 300 msec. After this period of adaptation, subjects were asked to respond to 5 stimuli as in the identification tests. Each stimulus in the series being tested was identified 8 times in randomized order in each condition by each subject.

RESULTS AND DISCUSSION

Although 7 subjects were consistently able to identify prevoiced labial stimuli among the synthetic stimuli presented, one subject never categorized any of the labial stimuli as prevoiced stops. The synthetic stimuli used in this experiment do not contain all the information available in natural speech for identifying prevoiced stops. Presumably this subject puts greater weight on other factors when judging whether a token is a prevoiced or a voiceless unaspirated stop. Since the design of this experiment assumes identification of three labial stops based on VOT variation, the data obtained from this subject were eliminated from further analysis.

Boundary points for each subject and for each condition were calculated using probit analysis. These data are presented in Tables 4 and 5. Significant boundary shifts were obtained for each condition except for adaptation of the labial series with the -70 msec labial stimulus ($p < .01$, by a t-test for correlated means). Table 6 outlines these results in the same format in which the predictions were presented.

As predicted by both hypotheses, the -70 msec velar adaptor did produce a shift on the velar series, while the -70 msec labial adaptor did not produce a shift on the labial series (cf. Donald, 1976; Foreit, 1977). Both of the adaptors with coincident voice onset shifted the voiceless inaspirate-voiceless aspirate boundary, as predicted by both hypotheses. The two cross-series conditions provide partial support for both of the hypotheses outlined. The -70 msec velar adaptor produced the shift on the labial boundary that was predicted by the phonetic hypothesis. The -70 msec labial adaptor produced a shift on the velar boundary as predicted by the auditory hypothesis.

A comparison of the magnitude of the boundary shifts yields the relative differences shown in Table 7. These differences are significant ($p < .05$, by a t-test for correlated means). As predicted by the auditory hypothesis, the 5 msec velar adaptor produced a larger boundary shift on the velar series than did either the -70 msec velar adaptor or the -70 msec labial adaptor. Furthermore, the 5 msec labial adaptor produced a larger effect on the labial series than did the -70 msec velar adaptor.

TABLE 4: Unadapted boundary points for velar voiceless inaspirate-voiceless aspirate boundary for Thai subjects on velar series, and boundary shifts.

	Unadapted	-70 msec velar adaptor *	5 msec velar adaptor *	-70 msec labial adaptor *
T.S.1	37.7	8.0	13.7	5.2
T.S.2	33.5	6.7	7.0	4.0
T.S.3	33.1	7.3	8.0	3.0
T.S.4	33.7	8.7	9.7	2.4
T.S.5	36.4	10.2	11.3	7.4
T.S.6	31.8	7.4	9.2	4.5
T.S.7	35.5	6.3	7.8	5.1
$\bar{x}$=	34.5	7.8	9.5	4.5

*denotes significant boundary shift, $p < .01$.

TABLE 5: Unadapted boundary points for labial voiceless inaspirate-voiceless aspirate boundary for Thai subjects on labial series, and boundary shifts.

	Unadapted	-70 msec labial adaptor *	5 msec labial adaptor *	-70 msec velar adaptor *
T.S.1	21.0	-.1	2.9	5.8
T.S.2	25.7	1.6	8.2	7.5
T.S.3	20.2	-1.1	3.3	1.8
T.S.4	22.6	2.1	5.1	4.3
T.S.5	20.2	-1.5	2.7	2.6
T.S.6	20.8	.5	4.5	6.4
T.S.7	26.6	1.5	8.5	9.0
$\bar{x}$=	22.4	.4	5.0	5.3

*denotes significant boundary shift, $p < .01$.

TABLE 6: Significant boundary shifts between voiceless inaspirates and voiceless aspirates obtained.

	Velar series	Labial series
-70 msec velar adaptor	shift	shift
-70 msec labial adaptor	shift	no shift

TABLE 7: Relative magnitude of boundary shifts obtained between voiceless inaspirates and voiceless aspirates ($p < .05$).

Velar series:	-70 msec velar adaptor	<	5 msec velar adaptor	> -70 msec labial adaptor
Labial series:	5 msec labial adaptor	>	-70 msec velar adaptor	

These results can be explained in terms of adaptation at both auditory and phonetic levels of processing. Adaptation of voicing at an auditory level has been attested in previous studies (Cooper, 1974; Tartter and Eimas, 1975). In the present experiment, adaptation of the velar boundary with the labial adaptor clearly is evidence for such auditory adaptation. Furthermore, the boundary shifts differ in magnitude as predicted by an auditory account. However, the presence of a boundary shift on the labial series produced by the -70 msec velar adaptor cannot be explained by an auditory process of the type proposed above, and this points to an additional contribution of phonological factors to the total effect.

REFERENCES

Bailey, P. (1973) Perceptual adaptation for acoustical features in speech. Speech Perception: Report on Research in Progress in the Department of Psychology. (Northern Ireland: The Queen's University of Belfast) 2, 29-34.

Cooper, W. E. (1974) Contingent feature analysis in speech perception. Percept. Psychophys. 16, 201-204.

Donald, S. L. (1976) The effects of selective adaptation on voicing in Thai and English. Haskins Laboratories Status Report on Speech Research SR-47, 129-136.

Foreit, K. G. (1977) Linguistic relativism and selective adaptation for speech: A comparative study of English and Thai. Percept. Psychophys. 21, 347-351.

Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague 1967. (Prague: Academia), 563-567.

McNabb, S. (1975) Must the output of the phonetic detector be binary? Research on Speech Perception, Progress Report 2. (Indiana University, Bloomington), 166-179.

Miller, J. L. (1976) Properties of feature detectors for speech: Evidence from the effects of selective adaptation on dichotic listening. Percept. Psychophys. 18, 389-397.

Tartter, V. C. and P. D. Eimas. (1975) The role of auditory feature detectors in the perception of speech. Percept. Psychophys. 18, 293-298.

Hemispheric Specialization for Speech Perception in Kindergarten Children with Language Deficiency

Davida R. Rosenblum[†] and Michael F. Dorman[††]

ABSTRACT

Twenty right-handed kindergarten children with superior language skills and twenty with deficient language, as defined by performance on an elicited sentence repetition task, were tested for hemispheric specialization for speech perception with a dichotic consonant vowel (CV) syllable task, and for relative manual proficiency by means of a battery of hand tasks. Reading readiness and aspects of other cognitive abilities were also assessed. The superior children evidenced a mean right-ear advantage (REA) of 14.5 percent, which is consistent with normal values reported by other investigators using the same stimuli. The language-deficient group evidenced essentially no mean ear advantage (0.5), with half of these subjects exhibiting left-ear superiority. The findings suggest relationships among cerebral dominance as inferred from dichotic testing, language proficiency (including reading readiness) and general cognitive functioning.

INTRODUCTION

Orton's theory of mixed dominance (Orton, 1937) associates speech, language and reading disorders with the failure of one hemisphere to dominate in the control of both motor and speech processes. Until recently, attempts

---

[†]Herbert H. Lehman College, City University of New York.

[††]Arizona State University, Tempe, Arizona.

Acknowledgment: This paper is based in large part on a dissertation submitted to the City University of New York by the first author, who wishes to acknowledge the invaluable contributions made by the second author, by Katherine S. Harris, Norma S. Rees and Michael Studdert-Kennedy, all of the City University. Thanks are also due M. Irene Stephens on extending to us the use of the experimental version of her elicited sentence repetition screening task, the staff of Haskins Laboratories for their assistance and helpful comments at various stages of the research, and the personnel of the New Rochelle school system, particularly Seymour Samuels, Beatrice Meisler and the speech staff for graciously providing access to their student body and facilities.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

to test this hypothesis have been hampered by the absence of reliable measures of handedness and by the absence of a nonintrusive method for assessing cerebral dominance for language. These obstacles have been partially overcome with the development of tests of relative manual proficiency (Satz, Achenbach and Fennell, 1967; Annett, 1970) and with the establishment of dichotic listening techniques as measures of hemispheric specialization for speech perception (Kimura, 1961b).

Orton's speculations linking various disorders with abnormal cortical lateralization have received, at best, equivocal support. Early dichotic studies suggested that stutterers evidenced abnormal lateralization (Curry and Gregory, 1969; Brady, Sommers and Moore, 1973). However, a recent study, using a dichotic nonsense-syllable task (Dorman and Porter, 1975), found no difference in lateralization for speech perception between normals and stutterers, and other studies using different dichotic tasks have reported similar results (Slorach and Noehr, 1973; Sussman and MacNeilage, 1975).

Variability of outcome also characterizes the reports of lateralization for speech perception in learning-disabled populations. Several studies have shown learning-disabled children to evidence abnormal lateralization (Kimura, 1963, for boys only; Zurif and Carson, 1970; Witelson and Rabinovitch, 1972; Dermody, Noffsinger, Hawkins and Jones, 1975). Once again, however, the use of a dichotic nonsense-syllable task has revealed no difference in lateralization for speech between dyslexic children and normal controls (Fischer, 1972). This outcome, in turn, is similar to those reported by Bryden (1970), Satz (1975), Witelson (1977) and Yeni-Komshian, Isenberg and Goldberg (1974).

Language deficient children, the target population of the present study, have also been subjected to diverse dichotic tasks, again with mixed results. Pettit and Helms (1974) and Starkey (1974) reported the absence of a right-ear advantage in language disordered children. However, Sommers and Taylor (1972) and Tobey, Cullen and Fleisher (1976) found normal ear advantages in similar populations.

In the present study, we assessed hemispheric specialization for speech perception in an extensively tested group of language deficient children and normal controls with superior language skills. For each child, aspects of overall intellectual functioning, articulatory ability, language comprehension and performance, reading readiness, and visual-motor integration were assessed and then correlated with handedness and dichotic listening performance. The dichotic task, Shankweiler and Studdert-Kennedy's nonsense syllables (1975), has been used before with normal children (Orlando, 1971; Dorman and Geffner, 1974; Geffner and Dorman, 1976) and is known to yield a REA of between 9 and 15 percent. Moreover, the task has also been used with two abnormal populations (stutterers and learning-disabled children) with the outcome of no differences between the normal and abnormal populations. Since the nonsense syllable test has proved to be conservative, in that it does not readily admit differences between the normal and abnormal populations, it would be particularly striking if a difference were found in lateralization for speech perception between normal and language deficient children.

METHOD

Subjects

Twenty language superior children (10 male; 10 female; mean CA=69.9 months) and twenty language deficient children (10 male; 10 female; mean CA=68.5 months) were selected from a pool of approximately 600 public school kindergarteners by the administration of an elicited sentence repetition task, the Stephens Oral Language Screening Test (Stephens, 1974). The control group was composed of children who made less than two error points in both repetition and articulation; the experimental group was composed of children who earned an error score of 25 or more in repetition, with articulation scores ignored. All children were right-handed as defined by performing at least two of three tasks (throwing a ball, writing, cutting with scissors) with the right hand. All children had normal and equal hearing in both ears and scored at least at the 90 IQ level on the Peabody Picture Vocabulary Test (Dunn, 1965). No child was bilingual, known to be organically impaired, a twin, a stutterer, a kindergarten repeater, or from the lowest-economic level (Group 7, Hollingshead, 1965).

Dichotic Listening Task

Synthetic signals appropriate for the six English stop consonant-vowel syllables (/ba, da, ga, pa, ta, ka/) were generated on the Haskins Laboratories parallel resonance speech synthesizer. Under computer control, these six stimuli were recorded dichotically in a fully counterbalanced, randomized order onto magnetic tape. The resulting tape contained 60 stimulus pairs with each member of a pair occurring twice on each channel. The interpair interval was 4 sec, with a 10 sec interval occurring after every 10 pairs. The signals were reproduced on a Panasonic RS 296 tape deck and presented via matched and calibrated TDH 39 headphones. The outputs of the tape channels were equated to within 2 dB and monitored before each test session. The signal level was 81 dB SPL.

The listeners were familiarized with the synthetic speech signals by three binaural presentations of the six test syllables. The 60 dichotic pairs were then presented twice, separated by an interval of 15-20 minutes, during which handedness tasks were administered. The headphones were reversed on the second run to control for channel effects. Only one response was elicited for each stimulus pair.

For each subject, a REA score was computed by the index R-L/R+L x 100, where R (or L) is the number of syllables correctly reported from the right (or left) ear. An absolute ear advantage score (AEA) was derived by simply eliminating the sign preceding the ear/advantage score. This index estimates the strength of lateralization without regard to side of dominance.

Measures of Handedness[1]

Each subject was tested on the hand preference section of the Harris Test of Lateral Dominance (Harris, 1957), and for relative manual proficiency for

---

[1]For details of administration of these and other measures, see Rosenblum, 1976.

peg placement, stylus tapping, card dealing, scissors cutting, and strength of grip. For each subject, a dextrality index (DI) was computed for each task using the same formula as for the ear advantage, except for the scissors cutting where L-R/R+L x 100 was used. A manual dexterity score was obtained by summing the raw scores on all the hand tasks.

Measures of Cognitive and Other Development

The IQ of each child was assessed by administration of the Goodenough Draw-a-Person Test (Goodenough, 1926) and the Peabody Picture Vocabulary Test (Dunn, 1965). Articulatory ability was assessed by the Fisher-Logemann Test of Articulation Competence (Fisher-Logemann, 1971). Language was measured by the Boehm Test of Basic Concepts (Boehm, 1971), a Complexity Measure of Expressive Language (Wurtzel, Roth and Cairns, 1976) and the Developmental Language Comprehension Test (Weiner-Mayster, 1975).[2] Reading readiness was determined by the Murphy-Durrell Reading Readiness Analysis (Murphy and Durrell, 1965). The single measure of nonlinguistic ability was the Developmental Test of Visual-Motor Integration (Beery, 1967).

RESULTS

IQ Language, Reading and Visual-Motor Tasks

The mean scores on these tasks for both groups are shown in Table 1. In addition to their inferior performance on elicited sentence repetitions, which was the sole determinant for group placement, the language deficient group performed significantly less well on the following tasks: Goodenough Draw-a-Person and manual dexterity (both $p < 0.05$), and Peabody Picture Vocabulary Test, sentence comprehension, comprehension of basic language concepts, expressive language complexity and visual-motor integration (all $p < 0.01$). The two groups also differed in performance on all of the subtests of the Murphy-Durrell Reading Readiness Analysis. However, since these scores were reported as percentiles, significance levels were not determined. There were no sex differences found on any of these measures.

Handedness

The two groups were right-handed to an equal degree on all six handedness tasks. There was a sex difference found for scissors cutting, with the girls achieving a greater between-hand difference ($p < 0.05$).

Dichotic Listening

The REA of the language deficient group (mean = 0.5) differed significantly from that of the superior children (mean = 14.5) ($p < 0.01$). By-subject inspection reveals that while only two of the control group children were left-eared (10 percent), nine of the language deficient children were left-eared and one showed no preference (50 percent). (See Figure 1 for the distribution of both groups' ear advantages.) The AEA also differed signifi-

---

[2]Weiner-Mayster, L. (1975) Developmental Language Comprehension Test. City University of New York, unpublished.

TABLE 1: Means and t-tests on all nonlaterality variables.

	Experimental	Control	t	p
Criterion Variables				
Sentence Repetition[a]	32.70	00.85	-13.36	<0.01
Age[b]	68.45	69.90	1.48	n.s.
Test Variables				
Language Comprehension	26.45	34.95	6.71	<0.01
Comprehension of Basic Concepts	30.74	42.20	6.16	<0.01
Complexity of Expressive Language	425.99	610.73	5.90	<0.01
Articulation (error score)[a]	13.05	00.80	not tested	
Peabody IQ	105.30	119.80	4.17	<0.01
Goodenough IQ	107.60	121.25	2.27	<0.05
Visual-Motor Integration[b]	64.25	76.15	4.75	<0.01
Manual Dexterity	298.82	332.22	2.55	<0.05
Reading Readiness[c,d]				
Phoneme Identification	72	99	not tested	
Letter Naming	82	89	not tested	
Learning Rate	78	99	not tested	
Total Score	82	96	not tested	

[a]Articulation errors were not scored above 3 errors during the administration of this test. In order to qualify for the control group, children could make no more than two such errors. Articulation was not a criterion for experimental group placement.
[b]In months.
[c]Median percentiles in months.
[d]Only 16 subjects in the experimental group and 18 subjects in the control group were tested on this variable.

FIGURE 1

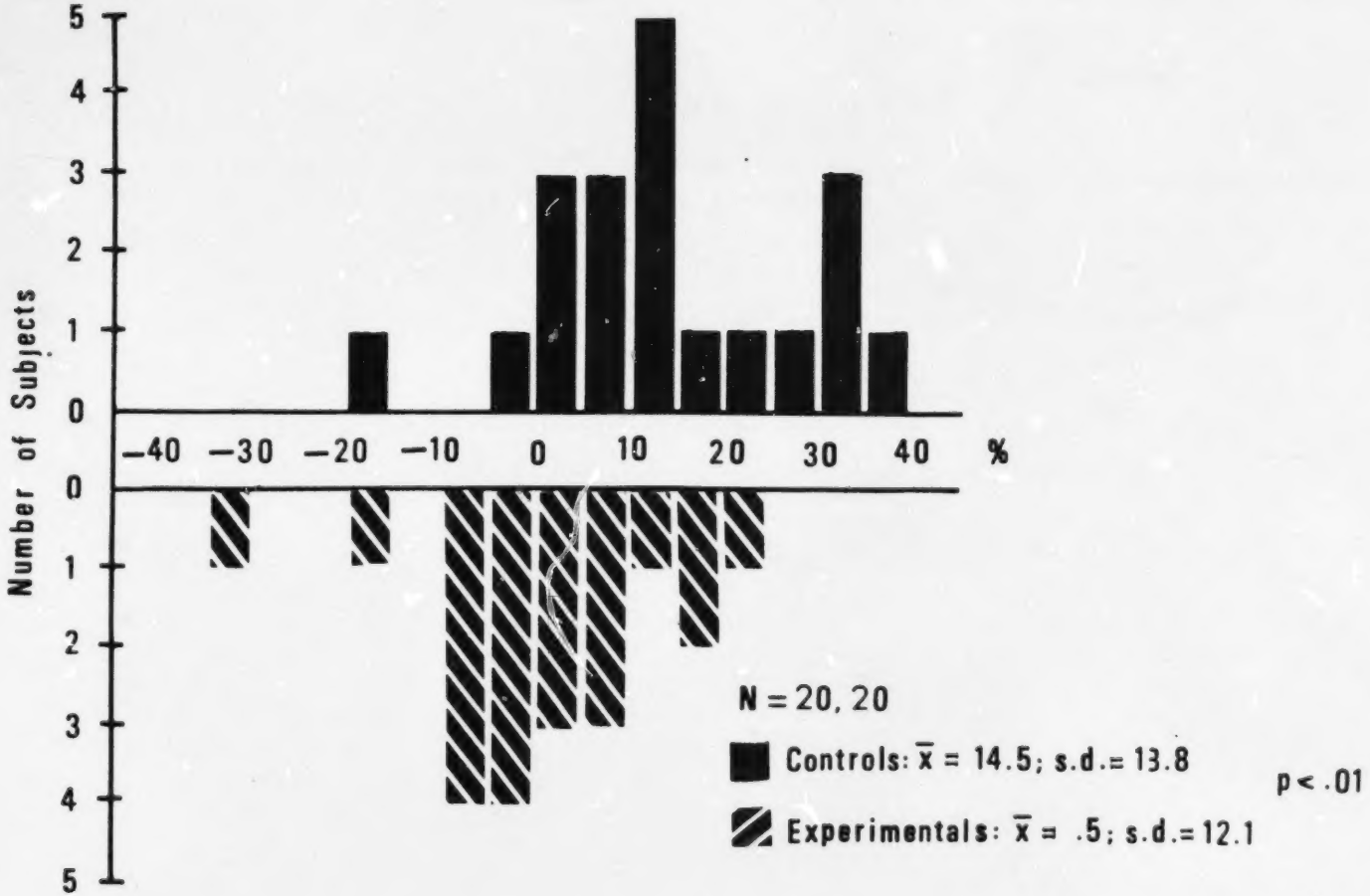

Figure 1: Distribution of the Right Ear Advantage.

cantly between the two groups (mean = 9.35 and 16.29 respectively, $p < 0.05$), indicating that the language deficient children were less lateralized even when direction of lateralization was ignored.

An error analysis of the dichotic listening results indicated that the two groups did not differ either in terms of the total number of errors, or the kind of errors (blend or place). The former outcome indicates that the absence of a REA in the language deficient group was not a floor effect due to poor overall performance on the task. In this respect, the language deficient children differed from developmental dyslexics (Witelson, 1977) who, although showing a normal REA, had fewer total correct responses than a normal control group. The language deficient children also performed differently than left-hemisphere damaged adults, who do not benefit from trials in which both members of the dichotic pair share the same place of articulation (double place cues) (Oscar-Berman, Zurif and Blumstein, 1975). In the present study both groups performed better with double place cues, but did not differ from each other in this respect. Only one sex difference was noted: the girls made somewhat more blend type errors than did the boys ($p < 0.05$).

IQ as a Factor in the REA

Since the two groups differed significantly in both Peabody IQ and Goodenough IQ, it is possible that the difference in lateralization between the two groups simply reflects the IQ difference. To determine whether lateralization for speech perception was related to IQ, the forty subjects were pooled and then divided into two new groups: those most strongly lateralized (mean AEA = 20.48) and those most weakly lateralized (mean AEA = 5.14). The difference in the resulting Peabody IQ means (strongly lateralized = 114.7; weakly lateralized = 110.4) was not significant. Therefore, Peabody IQ was independent of strength of lateralization. To assess in yet another manner the relationship between IQ and the ear advantage, two subsets of subjects from the two groups were matched on Peabody IQ (mean IQ = 116.5, n = 8) and the REAs examined. The REA of the language deficient group (mean = -5.47) remained significantly different from that of the superior group (mean = 18.45) ($p < 0.01$). Matching on the Goodenough IQ produced similar results (mean = 1.53 and 11.30 respectively, n = 8), although the difference failed to reach significance. In sum, the outcome of these several measures suggests that the REA difference between the two groups was not related to the IQ difference.

Simple Correlations

Intercorrelations among all the laterality measures are displayed in Table 2. For the normal group no correlations above 0.5 are found between the REA or AEA and any of the handedness tasks (this is an arbitrarily selected figure, based on the problems of determining a level of significance for a correlation matrix of the size constructed for this study, where measures were taken on only two samples).[3] In the language deficient group, however, a correlation of 0.71 between the REA and card dealing is shown. Among the handedness measures themselves, there is only one correlation of note, that

[3]Hayes, 1963, p. 576.

TABLE 2: Simple correlations among laterality measures.

Experimental Group

	Harris	Pegs	Tapping	Cards	Scissors	Grip	Absolute Ear Advantage
Right Ear Advantage	.21	.26	.02	.71	-.05	-.05	-.15
Harris		-.26	.33	.04	.02	.21	-.13
Pegs			-.21	.47	-.02	-.31	-.01
Tapping				.07	-.02	.36	-.01
Cards					.05	-.04	-.31
Scissors						.35	.17
Grip							-.09

Control Group

	Harris	Pegs	Tapping	Cards	Scissors	Grip	Absolute Ear Advantage
Right Ear Advantage	.12	.26	.19	-.19	.10	-.29	.84
Harris		.15	.04	-.45	.47	-.05	.23
Pegs			.63	.27	.10	.16	.31
Tapping				-.05	.18	.18	.22
Cards					-.24	.29	-.07
Scissors						-.17	.03
Grip							-.03

between stylus tapping and peg placement in the control group (0.63). Thus, the hand tasks appear to measure a set of unrelated skills that may be lateralized independently of one another. It also appears that no strong relations exist between the REA and any nonlaterality measure. Only a few correlations above 0.5 are found between the six sets of dextrality indices and the other measures; some of these are negative. Intercorrelations among the language measures ranged from low to moderate; however, discussion of these relationships will be dealt with in a subsequent article on the linguistic aspects of this study.

Stepwise Multiple Regression

Where generally positive but weak correlations exist, as was true in this sample, weighting the scores of several variables differentially may predict a dependent variable better than any one of them alone. A stepwise multiple regression method has been used previously for this purpose, with handedness measures combining to improve prediction of ear advantage [Orlando (1971) with eight- and ten-year old boys; Shankweiler and Studdert-Kennedy (1975) with adults]. In the present study, when the handedness measures were used as the predicting variables for REA, there were no significant increments above the simple correlation of the first variable used in the equation.[4]

Birth Order

Fifteen of the twenty language superior subjects were first-born or only children, whereas only three of the language deficient subjects held that position; the remainder of the experimental subjects were middle- or last-born.

DISCUSSION

Our results indicate that inability to repeat sentences accurately is associated with deviant ear asymmetries and a lower level of cognitive, linguistic and nonlinguistic functioning.[5] Elicited sentence repetition appears to reflect many of the components underlying children's language usage, including not only some facets of motivation and the ability to attend, but

---

[4]The handedness of children under the age of eight has been found to be extremely unreliable; it is believed that it has not yet achieved its full strength at the younger ages (Harris, 1957). This could explain the failure to improve ear/hand correlations in our sample with the multiple regression technique.

[5]It could be argued that the dichotic task reliability has not been established for these samples, and that there are two types of subjects whose side of ear advantage is likely to be reversed upon retesting; those who are left-eared and those who are weakly lateralized to either side (Blumstein, Goodglass and Tartter, 1975). Granted that this is the case, the fact remains that the experimental group in this study contains many more such subjects than does the control group, indicating that at the very least, their dominance cannot be determined as readily as can that of the control group children.

also short and long term auditory/linguistic memory. It is difficult then to isolate the factor or factors that may be responsible for poor performance on this task and for the accompanying atypical configuration of ear advantages found in these children; indeed, the weak correlations of sentence repetition scores and REA with the other measures suggests that the language deficient children in this study differ idiosyncratically in their areas of deficit.[6] It may be that the holistic nature of the elicited sentence imitation task is what makes it a good screening task: it will select children with a variety of deficits, any one or combination of which may be associated with abnormal lateralization.

Since the repetition task is not standardized, our use of the term "language deficient" is operational; we do not suggest that a score of 25 error points defines the borderline of deficient language performance. Although some subjects who are indeed language disordered in the clinical sense were probably selected by this measure, the relatively low error score chosen as the lower limit for inclusion in the "language deficient" group undoubtedly led to the selection of others who merely inhabit the lower end of the spectrum of normal language ability.[7]

Although both experimental and control groups were right-handed to an equal degree, the group of children who met our criterion of language deficiency evidenced virtually no ear advantage, while the language superior group evidenced a sizeable REA. An interpretation of these results must be tempered by at least two considerations: (1) there are many individuals in the general population who have no lateral dominance as measured by dichotic testing, yet who have normal or superior language skills, and (2) the differences found in the present study were between sample means--the distribution of scores overlapped. Thus, there were language deficient and language superior children who had similar ear advantages. These considerations indicate that the absence of a large REA does not necessarily imply abnormal language functioning.

Nevertheless, the difference between the mean REAs of our two groups was large and significant, and it is a striking fact that in dichotic studies of different right-handed populations, only neurologically impaired subjects have consistently exhibited such deviations from the expected magnitude and/or direction of the REA as were evidenced by the language deficient group of the

---

[6]Reliability may be low for many of our measures; retests could not be done, and for several of the tasks reliability has not yet been established. But even if reliability were found to be acceptably high in all cases, accurate prediction of sentence repetition scores could be expected only from those variables in which most of the children had deficits. There was no one area of which this was true.

[7]On three measures for which norms are available (Peabody IQ, Goodenough IQ and reading readiness), the language deficient group means did not fall below the norms even though they were substantially lower than those of the control group (see Table 1). In the case of the Visual-Motor Integration Test, the mean of the deficient group was six months below their chronological age, while that of the superior group was six months above.

study (Kimura, 1961a, 1961b; Goodglass, 1967; Curry, 1968; Schulhoff and Goodglass, 1969). In light of this, several hypotheses might account for some aspects of our data, although none is entirely satisfactory.

One hypothesis is based on the fact that birth stress has been implicated with anomalous dominance and language deficiencies (Bakan, Dibb and Reed, 1973; Kinsbourne, 1975). Right-handedness and no ear advantage (or left ear advantage) may then arise from covert lesion effects, and in these cases language development may be adversely affected.[8] According to this view all the dichotic scores in the experimental group would be interpreted as lesion effects. In other words, but for the postulated lesions, the degree of right-ear advantage in the right-eared children would have been greater, and the ear advantages of most of the left-eared children would have been shifted toward the right, bringing about a distribution of ear advantages that would approximate that of the control group.

However, this hypothesis has difficulty accounting for the fact that the language superior children were mostly only or first-born children, while the language deficient children were mostly middle- or last-born. The lesion model would have to suggest that middle or last-born children suffer greater prenatal or perinatal trauma than do first-born children, and it is by no means clear why this should be so.

An alternative hypothesis, consistent with the birth-order findings, would suggest that influences such as reduced linguistic input from adults, and different quality of input from siblings might account for the relatively poor linguistic skills of the experimental (later-born) groups of subjects. However, it is unlikely that such factors are so malevolent as to shift cerebral dominance.

A third hypothesis (Orton, 1937) suggests that the absence of a clear lateral preference arises from genetic mixing of right and left dominance. Once again, it is unclear why such intermixing should occur overwhelmingly in middle and last-born children.

In summary, the birth-order data appear incompatible with any of the three hypotheses. A resolution of this problem must await future dichotic studies in which the interactions among variables such as birth order, birth stress, major source of linguistic input during early childhood, handedness and cerebral dominance for language are thoroughly explored.

---

[8]It is interesting that the two mixed dominant (right-handed/left-eared) children in the language superior group earned scores below the median in more of the areas tested than did any one of the other superior subjects. This occurred for both children in manual dexterity, comprehension of basic concepts, sentence comprehension, syntactic complexity, Goodenough IQ, and number of dichotic errors, and for one or the other child in reading readiness, visual-motor integration and Peabody IQ. In contrast to this low overall pattern of achievement, each of the right-handed/right-eared children in the language superior group did very well in some areas and less well in others (for individual scores see Rosenblum, 1976).

REFERENCES

Annett, M. (1970) A classification of hand preference by association analysis. Brit. J. Psychol. 61, 303-321.

Bakan, P., G. Dibb and P. Reed. (1973) Handedness and birth stress. Neuropsychologia 11, 363-366.

Beery, K. E. (1967) Developmental Test of Visual-Motor Integration: Administration and Scoring Manual. (Chicago: Follett).

Blumstein, S., H. Goodglass and V. Tartter. (1975) The reliability of ear advantage in dichotic listening. Brain Lang. 2, 226-236.

Boehm, A. E. (1971) Boehm Test of Basic Concepts. (New York: The Psychological Corporation).

Brady, W., R. Sommers and W. Moore. (1973) Cerebral speech processing in stuttering children and adults. Paper presented at the convention of the American Speech and Hearing Association, Detroit, Michigan, November.

Bryden, M. P. (1970) Laterality effects in dichotic listening: Relations with handedness and reading ability in children. Neuropsychologia 8, 443-450.

Curry, F. K. W. (1968) A comparison of the performances of a right hemispherectomized subject and 25 normals on four dichotic listening tasks. Cortex 4, 144-153.

Curry, F. K. W. and H. Gregory. (1969) The performance of stutterers on dichotic listening tasks thought to reflect cerebral dominance. J. Sp. Hear. Res. 12, 73-82.

Dermody, P., P. D. Noffsinger, C. Hawkins and J. L. Jones. (1975) Auditory processing difficulties in children with learning disabilities. Paper presented at the convention of the American Speech and Hearing Association, Washington, D.C., November.

Dorman, M. and D. Geffner. (1974) Hemispheric specialization for speech perception in six year old children from low and middle socioeconomic classes. Cortex 10, 177-185.

Dorman, M. F. and R. J. Porter, Jr. (1975) Hemispheric lateralization for speech perception in stutterers. Cortex 11, 181-185.

Dunn, L. M. (1965) Peabody Picture Vocabulary Test. (Minneapolis: American Guidance Service).

Fischer, F. William. (1972) An analysis of reversal errors in children with severe reading disability: the relationship to certain linguistic and perceptual factors. Unpublished Master's thesis, University of Connecticut.

Fisher, H. B. and J. A. Logemann. (1971) The Fisher-Logemann Test of Articulation Competence. (Boston: Houghton Mifflin).

Geffner, D. and M. F. Dorman. (1976) Hemispheric specialization for speech perception in four-year old children from low and middle socio-economic classes. Cortex 12, 71-73.

Geschwind, N. and W. Levitsky. (1968) Human brain, left-right asymmetries in temporal speech regions. Science 161, 186-187.

Goodenough, F. L. (1926) Measurement of Intelligence by Drawings. (New York: Harcourt Brace and World).

Goodglass, H. (1967) Binaural digit presentation and early lateral brain damage. Cortex 3, 295-306.

Harris, A. J. (1957) The Harris Test of Lateral Dominance. (New York: Psychological Corporation).

Hayes, W. L. (1963) Statistics. (New York: Holt, Rinehart and Winston).

Hécaen, H. (1976) Acquired aphasia in children and the ontogenesis of

hemispheric functional specialization. Brain Lang. 3, 114-134.
Hollingshead, A. B. (1965) Two Factor Index of Social Position. (New Haven: Yale Station).
Kimura, D. (1961a) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.
Kimura, D. (1961b) Some effects of temporal lobe damage on auditory perception. Canad. J. Psychiat. 15, 156-165.
Kimura, D. (1963) Speech lateralization in young children as determined by an auditory test. J. Comp. Physiol. Psychol. 56, 899-902.
Kinsbourne, M. (1975) The ontogeny of cerebral dominance. In Developmental Psycholinguistics and Communication Disorders, ed. by D. Aaronson and R. W. Rieber. (New York: The New York Academy of Sciences).
Murphy, H. A. and D. D. Durrell. (1965) Murphy-Durrell Reading Readiness Analysis. (New York: Harcourt, Brace and World).
Orlando, C. (1971) Relationships between language laterality and handedness in eight and ten year old boys. Unpublished doctoral dissertation, University of Connecticut.
Orton, S. (1937) Reading, Writing and Speech Problems in Children. (New York: W. W. Norton).
Oscar-Berman, M., E. B. Zurif and S. Blumstein. (1975) Effects of unilateral brain damage on the processing of speech sounds. Brain Lang. 2, 345-355.
Petitt, J. M. and S. B. Helms. (1974) Cerebral dominance of language and articulation disordered children as measured by dichotic listening tasks. Paper presented at the convention of the American Speech and Hearing Association, Las Vegas, Nevada, November.
Rosenblum, D. R. (1976) Hemispheric specialization for speech perception in kindergarten children with language deficiency. Unpublished doctoral dissertation, City University of New York.
Satz, P. (1975) Cerebral dominance and reading disability: An old problem revisited. In The Neuropsychology of Learning Disorders, ed. by R. Knights and D. J. Bakker, Proceedings of NATO Conference. (Baltimore: University Park Press).
Satz, P., K. Achenbach and E. Fennell. (1967) Correlations between assessed manual laterality and predicted speech laterality in a normal population. Neuropsychologia 5, 295-310.
Schulhoff, C. and H. Goodglass. (1969) Dichotic listening, side of brain injury and cerebral dominance. Neuropsychologia 7, 149-160.
Shankweiler, D. and M. Studdert-Kennedy. (1975) A continuum of lateralization for speech perception? Brain Lang. 2, 212-225.
Slorach, N. and B. Noehr. (1973) Dichotic listening in stuttering and dyslalic children. Cortex 9, 295-300.
Sommers, R. K. and M. L. Taylor. (1972) Cerebral speech dominance in language-disordered and normal children. Cortex 8, 224-232.
Starkey, K. (1974) A dichotic test for subjects having limited functional speech and writing abilities. Paper presented at the convention of the American Speech and Hearing Association, Las Vegas, Nevada, November.
Stephens, I. (1974) The Stephens Oral Language Screening Test, experimental form. Unpublished test. (Lafayette, Indiana: Purdue University).
Sussman, H. and P. MacNeilage. (1975) Studies of hemispheric specialization for speech production. Brain Lang. 2, 131-151.
Tobey, E. A., J. K. Cullen and A. Fleisher. (1976) Performance of children with auditory processing disorders on a dichotic stop-vowel identification task. Paper presented at the convention of the American Speech and Hearing Association, Houston, Texas, November.

Wada, J., R. Clark and A. Hamm. (1975) Cerebral hemispheric asymmetry in humans. Arch. Neurol. 32, 239-246.

Witelson, S. F. (1977) Developmental dyslexia: Two right hemisphere and none left. Science 195, 309-311.

Witelson, S. F. and M. S. Rabinovitch. (1972) Hemispheric speech lateralization in children with auditory-linguistic defects. Cortex 8, 412-426.

Wurtzel, S., F. Roth and H. Cairns. (1976) A Complexity Measure of Expressive Language, Working Papers in Speech and Hearing Sciences, vol. 2. (Graduate School and University Center of the City University of New York).

Yeni-Komshian, G. H., H. Isenberg and H. Goldberg. (1974) Cerebral dominance and reading disability: left visual field deficit in poor readers. Neuropsychologia 2, 1-11.

Zurif, E. B. and G. Carson. (1970) Dyslexia in relation to cerebral dominance and temporal analysis. Neuropsychologia 8, 351-361.

Can the Intrinsic F_0 Differences Between Vowels Be Explained by Source/Tract Coupling*

William G. Ewan†

ABSTRACT

The fundamental frequency of voice (F_0) of vowels shows a slight systematic variation as a function of vowel height such that high vowels have the highest intrinsic F_0 and low vowels have the lowest intrinsic F_0. A simple experiment using [m] in the utterances [umu] and [ama] shows that these differences in F_0 cannot be explained by acoustic interaction (that is, the acoustic "pull" of F_1 on F_0). Although the formant structure should not differ significantly for either [m], the F_0 was significantly different. The F_0 of each [m] did not differ significantly from the F_0 of the following vowel. The results indicate that the intrinsic F_0 differences are caused by physical interaction between the vocal folds and supraglottal articulators which may alter the vocal fold mass or tension via interior soft tissue attachments.

The fundamental frequency of voice (F_0) varies directly with vowel "height." The F_0 of a high vowel such as [u] may be 5 to 25 Hz higher than the F_0 of a low vowel such as [a] in the same phonetic environment (see reviews of the literature by Atkinson, 1973 and Ohala, 1973). If the cause of this phenomenon were known, it might help us to understand the mechanism available to speakers for voluntary F_0 regulation.

A theory has been offered by Atkinson (1973) to account for the intrinsic variation in F_0 between vowels. In referring to Flanagan and Landgraf (1968), he suggested that the closer the first resonance (F_1) of the vocal tract is to the fundamental frequency of the vocal cords (F_0), the greater effect that resonance will have in raising F_0. This source/tract coupling would be greater for higher vowels since the F_1 is lower (and therefore closer to F_0) than it is for low vowels. This theory was challenged by Ohala (1973), with, among other things, the evidence (from Beil, 1962) that during helium speech, when all vowel formants are higher, a similar intrinsic F_0 difference between vowels still occurs and has the same magnitude as that during nonhelium speech.

---

*This is a "data expanded" version of a section of a Ph.D. Dissertation submitted to the University of California, Berkeley.

†Also, University of Connecticut Health Center, Farmington.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

Ewan (1976) offered another piece of evidence that did not support the "acoustic coupling" theory. The acoustic coupling theory should predict that the F_0 of a nasal stop, for example [m], should not show a change in F_0 due to a following vowel, since the lowest nasal resonance should not change significantly whether produced before [i], [a] or [u] (House, 1957; Fujimura, 1962). The nasal stop should eliminate any significant acoustic interaction created by the adjacent vowels and thus should "allow" the F_0 of the nasal stop to move to a "noncoupled" frequency. However, for the one subject in the study (Ewan, 1976), the F_0 of intervocalic nasal stops were similar to the F_0 of the following vowels (that is, [i], [a] and [u]) and significantly different from the F_0 of the same nasal stop preceding another vowel.

In order to confirm this finding, a simplified version of the experiment was performed at Haskins Laboratories. Ten naive subjects (5 male, 5 female) produced the two utterances [ama] and [umu] in the sentence frame: "Say _____ again." Twenty-one tokens of each type were produced with a level F_0 (in all but one case). A randomized list of the two utterances was used. The recordings were processed using the supervisor program of the PDP 11/45 computer and the GT-40 graphics terminal that samples the speech wave at a rate of 10 kHz. Each token was low-pass filtered at 600 Hz and then input to the computer and displayed. A manually controlled cursor was moved on the display to a point on the speech wave signal in the center of the nasal stop. The cursor was then moved to the end of that cycle. The distance between the first and last cursor points on the display (that is, one pitch period) appeared on the corner of the display in milliseconds. The utterance-final vowel was measured in the same way at a point approximately one-third of the way into the vowel. All measurements were recorded by hand.

The means for the two nasal stops (that is, $[m]_a$ and $[m]_u$) and the final vowels are shown in Table 1. A two-way analysis of variance was used with the pooled data from both male and female subjects to find whether the two nasal stops ($[m]_a$ and $[m]_u$) differed significantly in F_0 and whether the F_0 of $[m]_a$ and $[m]_u$ differed significantly from [a] and [u] respectively. The F_0 of $[m]_a$ and $[m]_u$ were significantly different ($p << .01$), while the F_0 of $[m]_a$ and [a], and $[m]_u$ and [u] did not differ significantly.

TABLE 1: Mean F_0 of [m] and the final vowels in the utterances [ama] and [umu].

	Males		Females		Combined	
	[m]	V	[m]	V	[m]	V
[u]	109.9	109.6	200.8	199.2	142.0	141.4
[a]	103.5	105.3	191.6	190.1	134.4	135.5

No evidence of source/tract coupling was found in this limited subject population. These results do not support the acoustic coupling theory of intrinsic vowel F_0. The nasal stop had an F_0 similar to that of the following vowel. Thus, it is quite reasonable to assume that intrinsic vowel F_0 effects are due to coarticulatory anticipation of the tongue, pharynx, or jaw for the following vowel. It is not clear what the role of each might be in altering F_0, since there is evidence that all three may be capable of altering F_0 through muscular or soft tissue connections that may alter the tension, vibrating mass, length or degree of approximation of the vocal folds [see relevant discussions in Ohala (1973) and Ewan (1976)][1].

REFERENCES

Atkinson, J. (1973) Aspects of intonation in speech: implications from an experimental study of fundamental frequency. Ph.D. dissertation. University of Connecticut, Storrs.

Beil, R. (1962) Frequency analysis of vowels produced in a helium-rich atmosphere. J. Acoust. Soc. Am. 34, 347-349.

Ewan, W. G. (1976) Laryngeal behavior in speech. Ph.D. dissertation. University of California, Berkeley.

Flanagan, J. L. and L. Landgraf. (1968) Self-oscillating source for vocal tract synthesizers. IEEE Trans. Audio Electroacoust. 16, 57-64.

Fujimura, O. (1962) Analysis of nasal consonants. J. Acoust. Soc. Am. 34, 1865-1875.

House, A. (1957) Analog studies of nasal consonants. J. Sp. Hear. Dis. 22, 190-204.

Ohala, J. (1973) Explanations for the intrinsic pitch of vowels. Monthly Internal Memorandum, Phonology Laboratory, University of California, Berkeley, January, 9-26.

[1]Ewan, W. G. Speech and laryngeal behavior. Report of the Phonology Laboratory, Phonology Laboratory, University of California, Berkeley (in preparation).

On the Relationship Between Vowel and Consonant Identification When Cued by the Same Acoustic Information

Paul Mermelstein†

ABSTRACT

When listening to speech, do we recognize syllables or phonemes? Information concerning the organization of the decisions involved in identifying a syllable may be elicited by allowing separate phonetic decisions regarding the vowel and consonant constituents to be controlled by the same acoustic information and by looking for evidence of interaction between these decisions. The duration and first formant frequency of the steady-state vocalic segment in synthesized consonant-vowel-consonant (CVC) syllables were varied to result in responses of /bɛd/, /bæd/, /bɛt/ and /bæt/. The fact that the duration of the steady-state segment controls both decisions implies that that segment must be included in its entirety in the signal intervals on which the two decisions are based. For most subjects, no further significant interaction between the vocalic and consonantal decision is found beyond the fact that they are both affected by changes in the duration parameter. Two separate phonetic decisions based on overlapping ranges of the signal are adequate models, and feedback from the output of the phonetic decisions need not be explicitly introduced.

INTRODUCTION

When listening to speech, do we recognize syllables or phonemes? The question of minimal perceptual units in speech perception has been thoroughly investigated. The dependence of consonantal place of articulation on the succeeding vowel in consonant vowel (CV) syllables (Delattre, Liberman and Cooper, 1955) argues for the indivisibility of such units. Studies on the effect of the preceding vowel duration on voicing of the final consonant in CVC syllables (Raphael, 1972) lead to similar conclusions. Experiments on backward recognition masking (Pisoni, 1972; Massaro, 1974) offer further support by revealing that consonant recognition is incomplete until a significant part of the vowel has been heard by the listener. Liberman (1970) has interpreted these findings by viewing the syllables as linguistic units within

---

†Also Bell-Northern Research and INRS-Telecommunications, University of Quebec, Montreal, Canada.

Acknowledgment: Anne Fowler's help with the generation of stimulus tapes and collection of the data is much appreciated. Quentin Summerfield provided many helpful comments on this manuscript. This work was supported in part by the National Institute of Child Health and Human Development.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

which phones exist only by way of their constituent features. Features carrying information about any one phoneme are generally temporally distributed, and features signaling separate phonemes overlap significantly. If the interaction between phonetic units exists entirely at the acoustic level, one may establish the criteria for each phonetic decision of a listener in terms of the acoustic variables alone. To succeed in this effort, one requires that the decision processes giving rise to the distinct phonetic components of the syllables be independent. However, if there exists additional interaction between the phonetic processes, then each decision will be determined not only by the current value of the acoustic features but also by all other phonetic decisions based on those features. Strictly hierarchic models of language understanding cannot account for such interaction between separate decisions on the same level. As long as such interactions are weak, models with feedback can be useful explanatory tools because they allow the output of one phonetic decision to be used as the input to a second one. If strong interactions are found, the usefulness of viewing distinct phonetic decisions based on the same acoustic information as separate must be questioned. Under such conditions, it would be a more parsimonious approach to admit only the existence of higher level units and to deny the relevance of direct perception of phones.

This paper reports on an experiment in which the interaction of vocalic and consonantal decisions was explored in a linguistic environment where the same set of acoustic parameters determined both decisions. We synthesized C_1VC_2 syllables where the spectral and temporal properties of the steady-state vocalic segment controlled the identity of V as well as the voicing feature of C_2. The boundaries in spectral-temporal space follow similar curves for most listeners, despite the fact that the 50 percent response lines may be significantly displaced from each other. Since the same information, the duration of the steady-state vowel segment, affects both the vocalic and the consonantal decisions, it was of interest to explore the interaction, if any, between these two decisions. Short durations result in a predominance of responses comprising the shorter vowel and an unvoiced stop. At longer durations, the syllable comprising the longer vowel and the voiced stop predominates. However, at intermediate duration values where the information regarding both the vowel and the consonant is ambiguous, no clear interaction between the two decisions could be found.

The boundaries for vowel perception in spectral-temporal space are of interest themselves, aside from any interactions between separate phonetic decisions. They allow a reexamination of previous suggestions based on speech-production data, that the listener takes into account the dynamic aspects of the speakers' production to interpret the spectral information (Lindblom, 1963). Active correction processes have been suggested (Halle and Stevens, 1962) as possibly mediating the task in an attempt to account for perceptual invariance of vowels in the face of significant spectral-temporal variations. Our results, insofar as durational variations of isolated CVC's allow one to make inferences concerning durational variations that may arise due to stress and rate, do not support the existence of such correction processes.

THEORY

The interdependence of the perceptual boundary between vowels along the spectral and temporal dimensions is of interest because it may shed light on the possible awareness of a listener regarding the speaker's production constraints. Lindblom (1963) has analyzed spectrographic data on the shift in vowel formants of speakers as a function of vowel length. He reported an "undershoot of the vowel target" at short durations. His interpretation of the data read as follows: "The talker does not adjust control of his vocal tract at fast rates to compensate for its response delay. His strategy of encoding is clearly not intended for a listener who demands absolute acoustic invariance in the realization of phonemes but it presupposes that the listener is able to correct for coarticulation effects." Such an interpretation is in line with the analysis-by-synthesis theory of speech perception (Halle and Stevens, 1962). The first-formant boundary between vowels can serve as a reliable indicator of any such correction. Existence of such active correction processes would predict an upward shift in first-formant frequency at the boundary with increasing vowel duration. A shorter vowel, on the assumption of insufficient time to reach the intended articulation, would be judged phonemically identical to a more open longer vowel that possesses a higher first-formant frequency.

That vowel duration affects the judgment of voicing in the postvocalic stop is well known (Denes, 1955; Raphael, 1972). The dependence of the same voicing feature on the spectral configuration of the vowel has received less attention. The more open the vowel, the greater change in F_1 is incurred in moving to the appropriate articulatory configuration from a stop configuration (Kent and Moll, 1969). Since longer vowel duration favors a voiced-consonantal percept, one expects to find more rapid spectral changes preceding an unvoiced stop. In prevocalic position a more rapid spectral change has, in fact, been found to favor the perception of a voiceless stop by Stevens and Klatt (1974). Subsequently, Summerfield and Haggard (1977) found that it was the higher F_1 onset frequency that acted as a positive cue for the perception of voiceless prevocalic stops. We may suspect, therefore, that if the postvocalic vowel-consonant transition is fixed in duration, but its starting F_1 frequency is increased, the likelihood of hearing a voiceless stop will also be increased. In turn, this could lead to an increase in the vowel-duration boundary that separates voiceless and voiced stops as F_1 of the vowel is increased.

The ability to control distinct phonemic decisions by varying the stimuli along the same acoustic continuum offers potentially interesting opportunities to explore the nature and interdependence of these decisions. As long as the distinct phonemic decisions are cued by a multiplicity of acoustic cues, any interaction found between the decisions may be due to the interaction of the acoustic cues themselves. One could argue that the listener extracts continuous but complex higher-order features from the simultaneous values of the distinct acoustic cues and arrives at phonemic decisions based on the values of these higher-order features. When the separate decisions are based on the same acoustic information and are found not to be independent, we are led to conclude that the outcome of one phonetic decision influences the outcome of a second such decision. If, on the contrary, these decisions appear to be independent, we are free to conclude that separate phonetic

decision processes adequately model the decoding of simple CVC stimuli.

Since we are concerned with the phonetic processing of syllable-sized stimuli, any interaction we find as a result of temporal variation of individual segments may provide evidence concerning an auditory segmentation of the stimulus prior to arriving at phonemic decisions. We may postulate a process that segments the stimulus in time somewhere in the vocalic region and assigns the vowel and consonantal categories on the basis of the duration of the resultant segments. Such a segmentation hypothesis would predict an interaction that increases the likelihood of a vowel decision appropriate to a longer vowel segment co-occurring with a consonantal decision appropriate to a shorter consonantal segment. Thus, if increasing the overall temporal duration increases the likelihood of hearing V_1 rather than V_2, and that of C_1 rather than C_2, the joint probabilities would be expected to follow the relation.

$$p(V_1, C_1) < (p(V_1) \cdot p(C_1), \text{ but } p(V_1, C_2) > p(V_1) \cdot p(C_2).$$

EXPERIMENT I - IDENTIFICATION OF VOWELS AND CONSONANTS

Method

Listeners were asked to identify synthetic speech stimuli as [bɛd], [bæd], [bɛt] or [bæt]. The /ɛ-æ/ vowel contrast was selected because of the strong effect of vowel duration on the identification of those vowels (Stevens, 1959). The /d-t/ consonant contrast was selected to exemplify the known dependence of the perception of voicing of a post-vocalic consonant on the vowel duration (Denes, 1955). The initial /b/ was added to all syllables so that they represented common English words.

The stimuli were prepared with the aid of a programmed (software) synthesis program with three adjustable formants. The program allowed the specification of acoustic segments in terms of duration, initial and final formant frequencies, formant frequency contours (linear, parabolic or cubic functions), voicing amplitude onset or offset and formant bandwidths. Starting with a series formant synthesis configuration, by addition of the individual formant bandwidth controls and overall-amplitude offset relative to the formant trajectory, the final transition was adjusted to give rise to voiced or voiceless stops depending on the vowel duration. The initial and final formant trajectories were linear with time and lasted 48 msec each. Initial and final formant frequencies corresponding to the stop articulations were set at 100, 1000, 2000 and 100, 2000 and 3000 Hz respectively. The variables controlled were the duration of the central stationary vocalic segment, D^V, and its first formant frequency, F_1^V. The second and third formants of the steady state segment were set at 1800 and 2500 Hz respectively. Two higher formants were fixed at 3500 and 4500 Hz. The formant bandwidths were set at 60, 80 and 100 Hz except for the final transition. There, the first formant bandwidth was adjusted to result in a first formant level whose drop-off increased linearly in magnitude to -20 dB over the extent of the transition. As a result of preliminary experiments, the F_1^V range was set to 625-675 Hz with 25 Hz increments. Ten distinct duration values were used for D^V, namely 48, 72, 80, 88, 96, 104, 112, 120,

160 and 240 msec. All duration values were integral multiples of the constant excitation duration of 8 msec, corresponding to a fundamental frequency of 125 Hz.

Stimuli were presented as identical pairs separated by a 0.5 sec pause. Five different randomizations of the set of 30 stimuli resulted in 150 presentations and responses by each listener. Eight listeners were selected, and all had at least some phonetic training so that they would be likely to make a consistent /ɛ/ vs. /æ/ vowel judgment.

Results

The vowel and consonant identification results of each subject were pooled separately and analyzed in the form of probability of identification as functions of duration at specified F_1^V values. Ogives were fitted to the frequency of /æ/ responses collapsed over both consonants and to the frequency of /d/ responses collapsed over both vowels.

The duration boundary values obtained by the above method are shown in Figure 1. The duration boundary for vowel identification is found to decrease with an increase of the F^V frequency for all subjects. The slope of the boundary ranges between 1 and 3 msec Hz. As expected, increasing F_1^V favors the likelihood of /æ/ responses, and there exists a trade-off relation between F_1 value and duration. The likelihood of hearing a relatively shorter but more open vowel as /æ/ can equal that of a relatively longer but less open vowel.

The variation in the duration boundary for consonant-voicing with changes in F^V is much less. Nevertheless, four out of the eight subjects showed an increase in the /t-d/ duration boundary for F^V values ranging between 625 and 675 Hz that is significant at the .05 level. For no subject can we find a significant decrease in the same boundary value. Evidently, an increase in F_1^V while the end-point of the final transition is kept fixed is a weak negative cue for the voicing feature of the postvocalic stop. As F_1^V is increased, the stop is more likely to be heard as voiceless, unless the D^V duration is also increased.

EXPERIMENT II - INTERACTION BETWEEN VOCALIC AND CONSONANTAL DECISIONS

Method

A second experiment was designed to probe more deeply into the interdependence between vocalic and consonantal decisions. The results shown in Figure 1 reveal a clear vowel-segment duration boundary in the perception of both the vowels and the postvocalic stops. For most subjects, a value can be found for the frequency of F_1 where the same duration value represents a perceptual boundary between the vowels and the consonants. Both the frequency and the duration values that together correspond to this condition vary from listener to listener. To probe the interdependence of these decisions in the region of maximum ambiguity, a simple adaptive procedure was designed that modified the parameters of each syllable-stimulus presented to the listener based on his previous syllable-identification response. The F_1^V frequency was modified based on the vowel response, the duration of the steady-state vowel

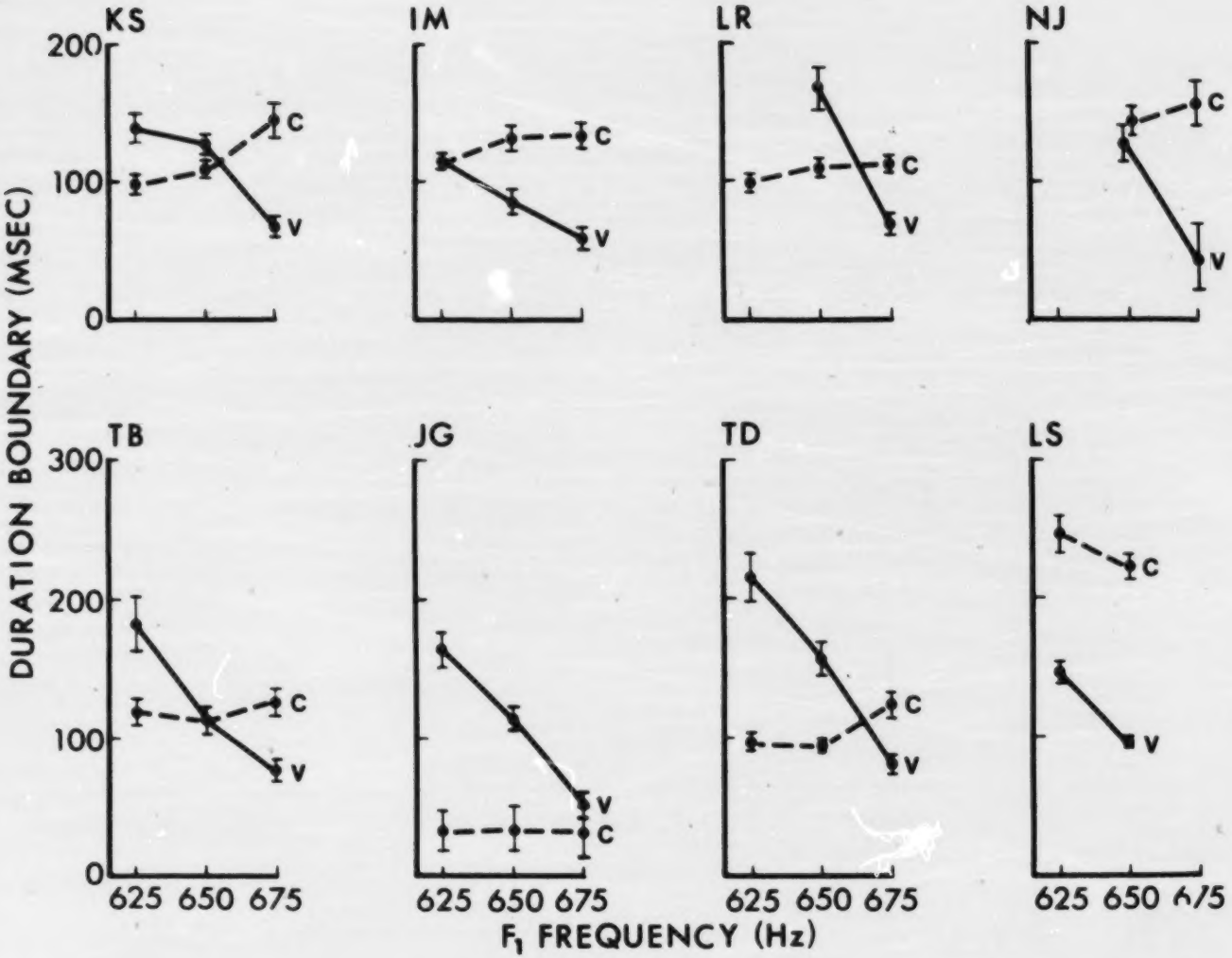

FIGURE 1

Figure 1: Variation of duration boundaries for vowel /ɛ-æ/ and consonant /t-d/ identification with F_1 of steady-state segments.

segment was modified based on the consonant-voicing response. The magnitude of the change in each parameter was constant, 10 Hz in F_1^V and 8 msec in D^V. Thus, a /bɛt/ response led to a +10 Hz change in F_1^V and a +8 msec change in D^V, while a /bæt/ response was followed by a -10 Hz change in F_1^V and a +8 msec change in D^V. The initial values for the parameters were set at 650 Hz and 104 msec respectively.

The experiment was carried out on-line at the computer, and listeners could ask for the same stimulus to be repeated as many times as they wished. Responses were indicated by keying the last two orthographic characters of the syllable at the terminal-keyboard. The experiment continued until 100 responses were collected from each subject.

Eleven subjects, all native speakers of English with no known hearing disabilities, participated in the experiment. Three subjects had previously participated in the off-line identification experiments, eight were new to the syllable identification task. Six subjects were unpaid volunteers from the laboratory staff, five subjects were students at Yale University and were paid $2.00 for their participation.

Results

The responses of ten subjects are shown in Table 1. Only responses to stimuli that were not perfectly consistently identified by the listener are included in the table. This restriction has the effect of excluding responses collected initially when the system adapted to the listener's ambiguity region as well as responses from the extremes of his ambiguity region where interaction due to duration cueing both vowel and consonant responses was more likely. One subject's results lacked sufficient consistency to allow an ambiguity region to be identified and his responses were eliminated from subsequent consideration.

TABLE 1: Summary of results on response interaction tests.

Subject	AT	ET	AD	ED	χ^2	Sig. at .05
AS	20	20	18	24	0.42	No
AF	23	18	24	25	0.45	No
GN	23	21	16	25	1.5	No
DK	18	18	16	19	0.13	No
TB	22	25	20	21	0.03	No
LR	11	28	30	15	-12.4	Yes
DW	12	16	15	16	0.18	No
JG	5	23	20	6	-18.9	Yes
AL	29	16	15	33	10.2	Yes
MS	30	15	17	31	9.1	Yes
Total	193	200	191	215	0.34	No

Typical results, such as those of subject AF, are shown in Figure 2. The responses from the outlined ambiguity region were collected and one interaction test was performed on the total responses of each syllable to the stimuli contained in that region. Such pooling of responses can be justified when the region of stimulus parameters is sufficiently small and insufficient responses are available to any one stimulus to obtain reliable separate interaction estimates. A one degree of freedom χ^2 test at the .05 level of significance yielded negative results for six subjects and oppositely directed positive results for two groups of two subjects. These results lead to a model of phonetic processing where the syllable response is made up of independent vowel and consonant decision processes operating on overlapping acoustic information. Figure 3 illustrates such a model.

Differences between the response patterns of individual subjects are large. Yet interaction pooled among all subjects is not significant. We have no information, as yet, on the sources that give rise to the divergence of results between individual subjects. Linguistic background, phonetic training and degree of attention paid to the experiment could be just some of the factors.

CONCLUSIONS

Vowel and consonant identification in synthesized syllables can both be controlled by variations in the same acoustic parameter, namely the duration of the steady-state vowel segment. The standard errors for the equal identification probability values are not significantly different in the two cases. It is therefore highly unlikely that different processes are invoked for duration discrimination for the purposes of identification of vowels and consonants as suggested by Pisoni (1971). When linguistic categories are imposed on both the vowel and consonant continua, identification as a function of duration is comparable.

Our results generally support a view of syllable perception where independent phonetic decision processes operate on overlapping segments of the acoustic signal. These results are in agreement with those of Huggins (1968) who also failed to find perceptual compensation in timing between vowel and consonant in the same syllable. Temporal segmentation of the syllable segment into vowel and consonant parts would predict a negative interaction between the probability of hearing a longer vowel and the probability of hearing a voiced stop. For only two of ten subjects is such correlation found at a significance level exceeding .05. For the other eight subjects no such interaction is found, therefore a segmentation model can be rejected as not generally appropriate for the interpretation of the results obtained above.

Draper and Haggard (1974), on the basis of similar interaction tests between the place and voicing features of the same consonant, suggest a feedback model between the various feature decisions required to establish consonant identity. In contrast, our results indicate an absence of significant interaction for most subjects between vowel and postvocalic voicing decisions. We must therefore conclude that in our experiments the decisions are carried out independently and without significant participation of feedback processes.

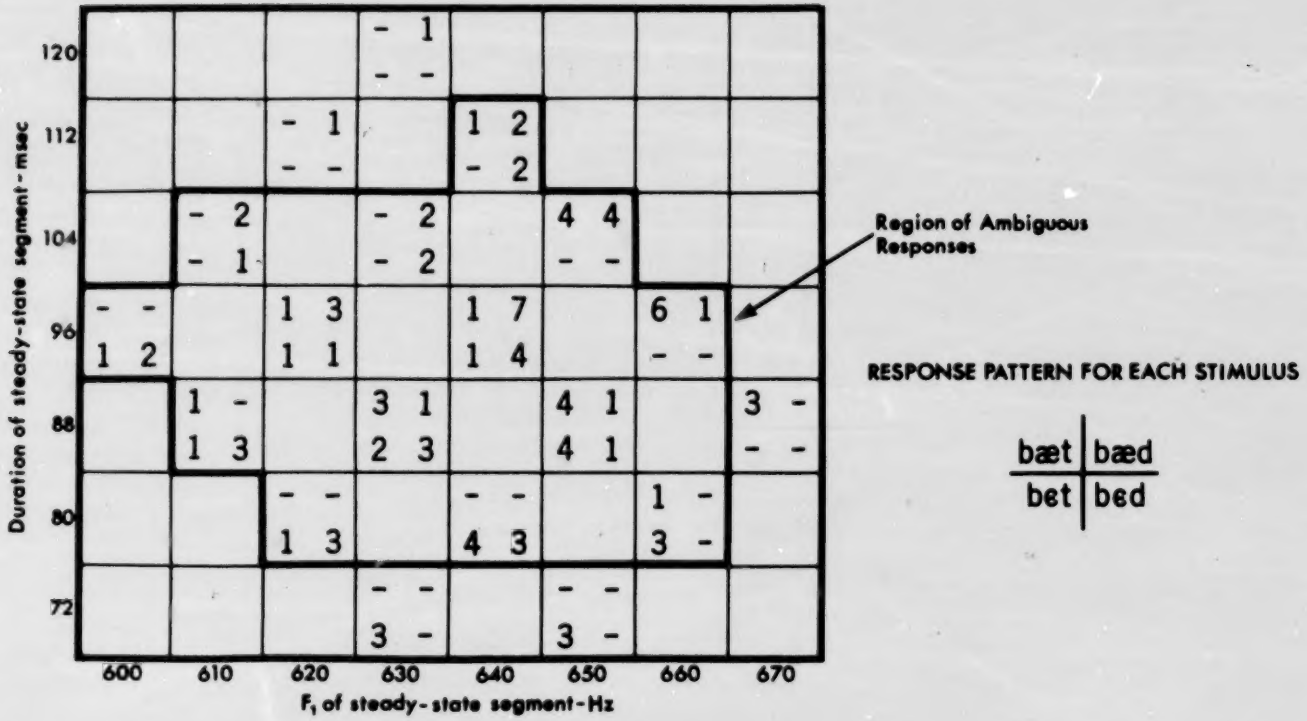

Figure 2: Response pattern for subject AF in interaction test.

FIGURE 3

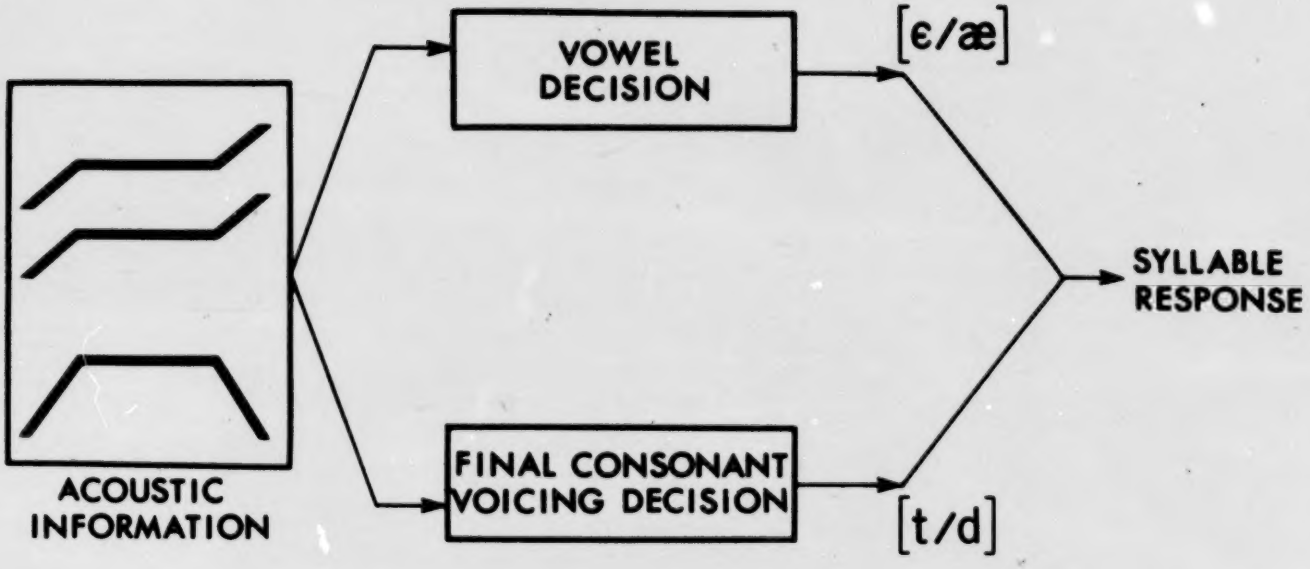

Figure 3: Model of vowel and consonant decisions operating on same acoustic signal to yield a syllable response.

Our results offer no support to an interpretation that a listener makes allowance for the temporal production constraints of the speaker as suggested by Lindblom (1963). Lindblom's suggestions would predict that increased duration acts as a negative cue for the more open vowel. If the listener assumes that the speaker's temporal constraints prevented him from reaching his intended target articulation, he should interpret a decrease in duration as a positive cue for the more open vowel. For all subjects the duration boundary for vowel identification showed contrary results. The spectral-temporal interaction is such that both longer duration and increased F_1 act as positive cues for the more open vowel. Similar results have been found for these vowels when presented not in the context of a CVC syllable but in isolation.[1]

The predictions based on temporal production constraints pertain more directly to talking rate than to duration differences. Lindblom (1963) used rate and stress to control duration in the data to be analyzed. Now, Fujisaki, Nakamura and Imoto (1975) show that talking rate influences the identification boundary for Japanese vowels when that boundary is based on duration. Admittedly, length and spectral differences may be more independent for Japanese vowels than for English vowels. Nevertheless, they find that listeners adjust their perceptual boundary in accordance with the talking rate, that is, the duration boundary is shorter at higher talking rates than at lower rates. Our results show that the duration boundary is also shorter at higher F_1 values. Therefore, our results would lead one to expect that at faster talking rates, which result in shorter durations, the F_1 vowel identification boundary would be higher than at slower rates. This is contrary to Lindblom's prediction. One might possibly argue that by presenting stimuli regularly spaced in time, our experimental conditions did not allow the listener to establish differential expectations concerning the synthesized vowel durations. However, transformation of increased duration from a positive cue for open vowels in isolation to a negative cue when these vowels are embedded in connected speech appears an unlikely possibility.

REFERENCES

Delattre, P. C., A. M. Liberman and F. S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.

Denes, P. (1955) Effect of duration on the perception of voicing. J. Acoust. Soc. Am. 27, 761-768.

Draper, G. and M. Haggard. (1974) Facts and artifacts in feature independence. Proc. Speech Communication Seminar, Stockholm, 197-219.

Fujisaki, H., K. Nakamura and T. Imoto. (1975) Auditory perception of duration of speech and nonspeech stimuli. In Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (London: Academic Press).

Halle, M. and K. N. Stevens. (1962) Speech recognition: A model and a program for research. IRE Trans. Info. Theory IT-8, 155-159.

Huggins, A. W. F. (1968) The perception of timing in natural speech I; compensation within the syllable. Lang. Speech 11, 1-11.

Kent, R. D. and K. L. Moll. (1969) Vocal-tract characteristics of the stop

[1]Mermelstein, unpublished.

cognates. J. Acoust. Soc. Am. 46, 1549-1555.
Liberman, A. M. (1970) The grammars of speech and language. Cog. Psych. 1, 301-323.
Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 38, 1773-1781.
Massaro, D. W. (1974) Perceptual units in speech perception. J. Exp. Psych. 102, 199-208.
Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Supplement to Haskins Laboratories Status Report on Speech Research.
Pisoni, D. B. (1972) Perceptual processing time for consonants and vowels. Haskins Laboratories Status Report on Speech Research SR-31/32, 83-92.
Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. J. Acoust. Soc. Am. 51, 1296-1303.
Stevens, K. N. (1959) Effect of duration upon vowel identification. J. Acoust. Soc. Am. 31, 109(A).
Stevens, K. N. and A. S. House. (1972) Speech Perception. In Foundations of Modern Auditory Theory, vol. 2, ed. by J. V. Tobias. (New York: Academic Press), 1-62.
Stevens, K. N. and D. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.
Summerfield, Q. and M. Haggard. (1977) On the dissociation of spectral and temporal cues to the voicing decision in initial stop consonants. J. Acoust. Soc. Am. 62, 435-448.

Information Conveyed by Vowels: A Confirmation*

David Dechovitz†

ABSTRACT

An early study with synthetic speech suggested that vowel identification includes a normalization stage, such that the listener calibrates his perceptual apparatus for each talker's vowel space (Ladefoged and Broadbent, 1957). In the present study, evidence for such a perceptual mechanism was obtained using natural speech. Each of a set of b-vowel-t test words, spoken rapidly within a sentence carrier by an adult male, was presented for recognition within a carrier selected from the adult's productions, appropriately embedded within an identical carrier produced by a nine-year-old male with substantially different vocal tract dimensions and excised from sentence context. The two talkers achieved the same pitch levels and speaking rate, with the result that mixed-talker sentences were perceived as if uttered by one talker. Errors in recognition were most substantial for test words embedded within the child's carrier. Apparently, the identity of a vowel may be computed over some short stretch of speech longer than the syllable in which it lies.

The acoustic structure of speech varies from one talker to another. For example, to the extent that formant frequencies reflect vocal tract dimensions, the absolute positions of the formants are not the same for a child as they are for an adult. Thus, current theories of the vowel stress the relational nature of the acoustic cues, since no absolute values of formant frequencies could unambiguously distinguish vowels produced by different talkers (cf. Shankweiler, Strange and Verbrugge, 1977).

The intrinsic variability of the acoustic structure of vowels has led some investigators to propose that the listener calibrates (normalizes) each talker's vowel space on the basis of some reference derived from preceding utterances (Joos, 1948; Ladefoged and Broadbent, 1957; Gerstman, 1968; Lieberman, 1973). There is experimental evidence from identification studies of synthetic vowel targets that has been taken as support for this hypothesis. Significant response variations to synthetic /b-V-t/ targets were shown to occur with certain variations in the overall frequencies of a precursor

---

*This paper was presented at the 93rd meeting of the Acoustical Society of America, 6-10 June, 1977, at the Pennsylvania State University, University Park, Pennsylvania.

†Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-51/52 (1977)]

phrase; listeners seemed to interpret vowels and precursors as if they had been produced by the same size vocal tract (Ladefoged and Broadbent, 1957). In another investigation, which used test syllables representing different vocal tracts, increases in reaction time for vowel identification were parsimoniously explained in terms of normalization (Summerfield and Haggard, 1973).

However, in the first study mentioned above, the vowel targets showed marked stability after some precursor phrases that differed considerably from other precursors in F_1/F_2 frequency ranges. Furthermore, prior exposure to any subset of a talker's vowel productions has not been shown to bring about a systematic reduction of errors in identifying target vowels (in syllabic context) uttered by that talker (Verbrugge, Strange, Shankweiler and Edman, 1976). There is also evidence that a consonantal environment does not aid in specifying a naturally produced vowel by providing a normalizing function (Strange, Verbrugge, Shankweiler and Edman, 1976). In fact, it seems that speculation on the vowel constancy problem has failed to give due emphasis to the richness of the natural speech signal. It is doubtful that first- and second-formant frequencies exhaust the sources of information that specify a vowel in natural productions (Miller, 1953; Peterson, 1961; Ladefoged, 1967). Other research, in fact, has emphasized the relative duration of preceding segments as contributing to the perception of vowels with short and long syllabic nuclei (Ainsworth, 1972). Moreover, a number of studies have revealed remarkably low error rates when listeners identify naturally produced isolated syllables spoken by voices with which they have no prior experience (Peterson and Barney, 1952; Abramson and Cooper, 1959).

Thus, there is currently no consensus about the perceptual problem posed by vowels uttered in the context of a single syllable, nor about the information acquired during experience with a voice. All studies producing changes in the identification of test vowels by varying vocal tract dimensions have been constructed with synthetic targets. It would be illuminating to verify such demonstrations with natural speech, in which all the potential sources of information ordinarily available to perceivers are present.

To provide verification, an adult male and a nine-year-old male individually recorded an identical set of sentences. Each talker, guided by a metronome, rapidly produced the sentence frame, "Please say ___ for me," several times at identical rates, embedding within it on different repetitions one of five test syllables: /bɪt/, /bɛt/, /bæt/, /bɑt/, or /bʌt/. The adult talker virtually duplicated the pitch levels, speaking rate, and stress pattern of the nine-year-old, thus confining the main differences between speakers to the relative formant frequencies.

The adult talker's /b_t/ productions were placed in three environments. One token of each /b_t/ target was (1) excised from sentence context, (2) embedded within a sentence frame selected from the adult talker's recordings, and (3) embedded within a selected frame from the nine-year-old's recordings. A different group of 25 listeners heard a 20-item randomized identification test for each experimental condition, each test containing four presentations of the five test syllables. Each test sequence was presented over a single loudspeaker.

The results are shown in Figure 1. The figure indicates an average 2.3 percent errors for identification of test syllables presented within a natural frame (that is, the adult's carrier sentence), 14.8 percent average errors for test syllables presented in isolation, and 54 percent average misidentification for two test syllables embedded within a spectrally inappropriate carrier (that is, the child's carrier sentence). The nature of the particular vowel confusions in each test environment suggest two ways in which factors beyond the syllable shape the perceptual specification of vowels.

An analysis of listeners' errors in identifying isolated test syllables reveals a response tendency toward the shorter vowel alternatives: that is, hearing /bɛt/ as /bɪt/, /bɑt/ as /bʌt/, and /bæt/ as /bʌt/, (although there was an equally strong tendency to hear /bæt/ as /bɑt/). It is also apparent, at least for /bɛt/ and /bɑt/, that perceptions migrated to shorter alternatives that were most spectrally similar to the vowels intended. In general, these response shifts suggest that listeners treated the isolated syllables (excised, it will be recalled, from a sentence frame) as if they had been spoken more slowly in citation form; listeners failed to preserve vowel identity in the acoustic transformations produced by rapid articulation. Thus, it seems that information about a talker's tempo is critical in achieving constancy, and that this information is not completely specified at the level of the single syllable.

In comparison, error rates for syllables in natural context were substantially lower; there were no response biases of the sort found for isolated syllables. A recent study reported similar differences between rapidly articulated destressed /p-vowel-p/ syllables presented in isolation and in the sentence context in which they were produced (Verbrugge, Shankweiler, Strange and Edman, 1976). Thus, a temporally appropriate carrier apparently contains sufficient information to accurately specify rate of articulation, and as a consequence, to reduce ambiguity in vowel perception.

The nature of errors for syllables embedded within the child's carrier frame suggests a second and more powerful manner in which extrasyllabic factors are perceptually salient for vowel identification. In particular, two vowel confusions accounted for most of the errors in this test environment: listeners misperceived 50 percent of the /bɛt/ presentations as /bʌt/, and 48.2 percent of the /bæt/ presentations as /bɑt/. A comparison of the F_1 and F_2 frequencies delimiting acoustic space for each talker (Figure 2) reveals the significance of these confusions. One sees that the adult's /bɛt/ and /bæt/ syllables are most spectrally similar to /bʌt/ and /bɑt/ of the child's productions. Thus, these results suggest that listeners adjusted their criteria for vowel identification according to formant ranges specified by contextual material; vowel identification was sensitive to the acoustic transformations produced by cross-vocal tract variation. Note that this account predicts shifts only for the /bɛt/ and /bæt/ syllables among the syllables used, since only these vowels of the adult are more similar to alternative vowels in the child's vowel space.

On the other hand, listeners accurately perceived intrinsic vowel duration despite the confusions produced by spectral characteristics of the carrier. The vowel /ɛ/, as articulated in the adult's /bɛt/ syllable, although similar in F_1 and F_2 frequencies to both /ɑ/ and /ʌ/ of the child's

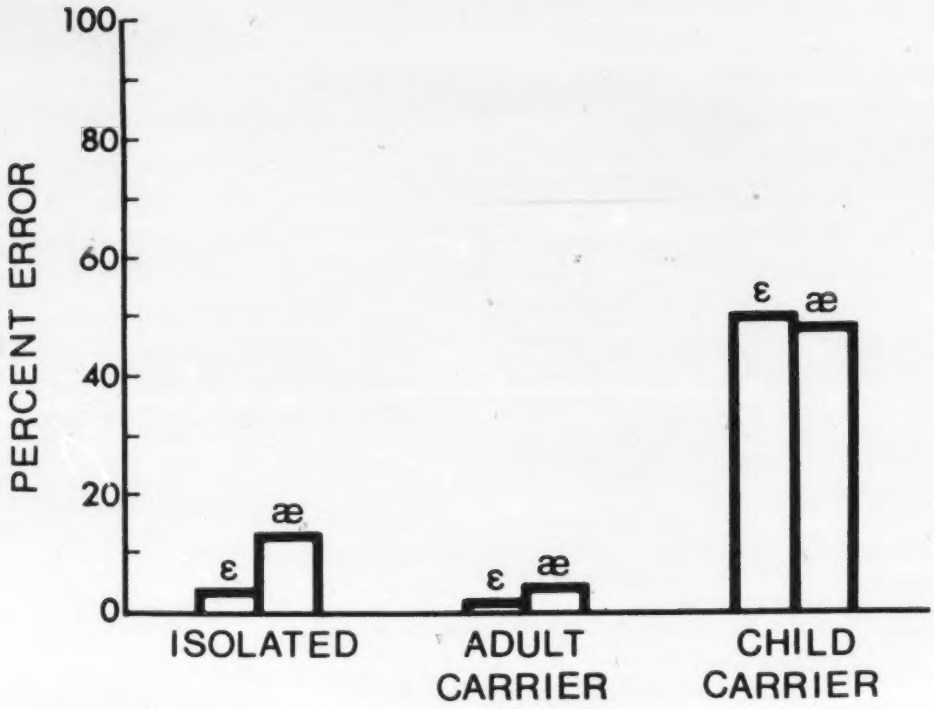

Figure 1: Average percent errors for identification of five test vowels in three environments.

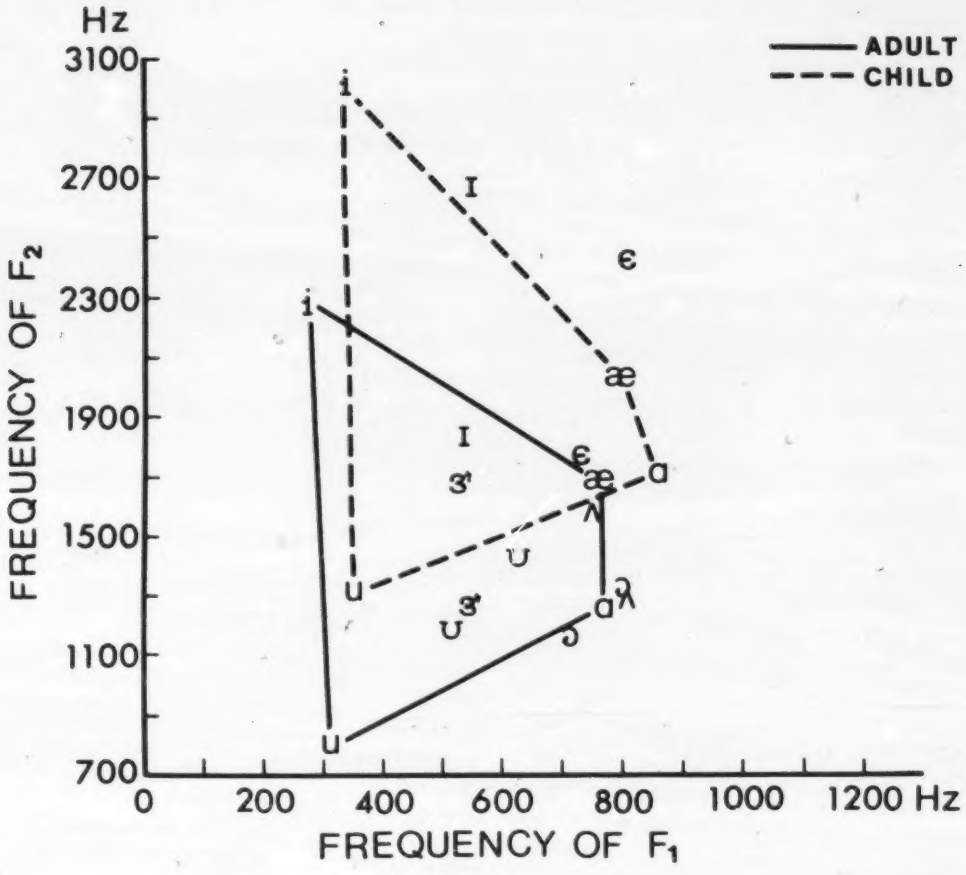

Figure 2: Comparison of the F_1 and F_2 frequencies delimiting acoustic space for the adult and child talkers.

productions, predominantly shifted to /ʌ/, to which it is more similar in duration. Similarly, the adult's /æ/ vowel, although spectrally resembling the child's /ɑ/ and /ʌ/, shifted to /ɑ/, to which it, in turn, is more similar in duration. Further, more generally, the pattern of errors in this test environment did not show a lack of compensation for the acoustic effects of rapid articulation. As suggested above, then, sentence context apparently aids vowel identification by permitting adjustment to a talker's tempo. It is also clear that a range of first- and second-formant frequencies does not exhaust the sources of information that specify a vowel in the natural speech signal.

We thus have further support for the view that vowels are relational values. The identity of a vowel may be computed over some short stretch of speech that is longer than the syllable in which it lies.

REFERENCES

Abramson, A. S. and F. S. Cooper. (1959) Perception of American English vowels in terms of a reference system. Haskins Laboratories Quarterly Progress Report QPR-32, Appendix 1.

Ainsworth, W. A. (1972) Duration as a cue in the recognition of synthetic vowels. J. Acoust. Soc. Am. 51, 648-651.

Gerstman, L. H. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Joos, M. A. (1948) Acoustic phonetics. Lang. (S)24, 1-136.

Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (New York: Oxford University Press).

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Am. 29, 98-104.

Lieberman, P. (1973) On the evolution of language: A unified view. Cog. 2, 59-94.

Miller, R. L. (1953) Auditory tests with synthetic vowels. J. Acoust. Soc. Am. 25, 114-121.

Peterson, G. E. (1961) Parameters of vowel quality. J. Sp. Hear. Res. 4, 10-29.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in the study of the vowels. J. Acoust. Soc. Am. 24, 175-184.

Shankweiler, D., W. Strange and R. Verbrugge. (1977) Speech and the problem of perceptual constancy. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N.J.: Erlbaum Assoc.)

Strange, W., R. Verbrugge, D. Shankweiler and T. Edman. (1976) Consonant environment specifies vowel identity. J. Acoust. Soc. Am. 60, 213-224. [Also Haskins Laboratories Status Report on Speech Research SR-45/46, 37-61.]

Summerfield, A. W. and M. P. Haggard. (1973) Vocal tract normalization as demonstrated by reaction times. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University of Belfast) Series 2, no. 3, 1-26.

Verbrugge, R., D. Shankweiler, W. Strange and T. Edman. (1976) Shifts in vowel perception as a function of speaking rate. Haskins Laboratories Status Report on Speech Research SR-47, 165-170.

Verbrugge, R., W. Strange, D. Shankweiler and T. Edman. (1976) What informa-

tion enables a listener to map a talker's vowel space? J. Acoust. Soc. Am. 60, 198-121. [Also Haskins Laboratories Status Report on Speech Research SR-45/46, 63-94.]

II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS

Abramson, Arthur S. (in press) The phonetic plausibility of the segmentation of tones in Thai phonology. Proc. XIIth Intl. Cong. Linguists, Vienna, 29 Aug. - 2 Sept., 1977.

_____. (1977) Laryngeal timing in consonant distinctions. Phonetica 34, 295-303.

Baer, Thomas. (1976) Effects of subglottal pressure changes on sustained phonation. J. Acoust. Soc. Am. 60, 365.

Crowder, Robert G. (in press) Audition and speech coding in short-term memory. In Attention and Performance, vol. 7. (Hillsdale, N.J.: Lawrence Erlbaum Associates).

_____. (in press) Language and memory. In Speech and Language in the Laboratory, School, and Clinic, ed. by J. F. Kavanagh and W. N. Strange. (Cambridge: M.I.T. Press).

_____. (in press) Memory for phonologically uniform lists. J. Verb. Learn. Verb. Behavior. 17.

_____. (in press) Sensory memory systems. In Handbook of Perception, vol. 9, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press).

Drewnowski, Adam and A. F. Healy. (in press) Detection errors on the and and: Evidence for reading units larger than the word. Memory and Cognition.

Freeman, Frances J., E. S. Sands and K. S. Harris. (in press) Temporal coordination of phonation and articulation in a case of verbal apraxia: A voice onset time study. Brain. Lang.

Gay, Thomas. (in press) Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am.

Healy, Alice. (1977) Pattern coding of spatial order information in short-term memory. J. Verb. Learn. Verb. Behav. 16, 419-437.

Kiame, M. Y.-S., M. C. Liberman and T. Baer. (1977) Tuning curves of auditory-nerve fibers. J. Acoust. Soc. Am. 61, S27.

Kubovy, Michael and A. F. Healy. (1977) The decision rule in probabilistic categorization: what it is and how it is learned. J. Exp. Psych.: Gen. 106, 427-446.

_____ and A. F. Healy. (1977) Numerical decision and the ideal learner: A reply to Dorfman. J. Exp. Psych.: Gen. 106, 450-452.

Lisker, Leigh. (1977) Factors in the maintenance and cessation of voicing. Phonetica 34, 304-306.

Niimi, Seiji, T. Baer and D. Fujimura. (1976) Stereo-fiberscopic investigation of the larynx. J. Acoust. Soc. Am. 60, 364.

Proffit, Dennis and T. Halwes. (in press) Categorical perception: A contractual approach. In Cognition and the Symbolic Processes, vol. 2. (Hillsdale, N.J.: Lawrence Erlbaum Associates).

Repp, Bruno H. (1977) Measuring laterality effects in dichotic listening. J. Acoust. Soc. Am. 62, 720-737.

_____. (in press) Stimulus dominance and ear dominance in the perception of dichotic voicing contrasts. Brain Lang.

_____. (1977) Perceptual interactions between implosive and explosive formant transitions of intervocalic stop consonants. Bull. Psychon. Soc. 10(A), 244.

Sands, Elaine, F. J. Freeman and K. S. Harris. (in press) Progressive changes in articulatory patterns in verbal apraxia: A longitudinal case

study. Brain Lang.
Sawashima, Masayuki and F. S. Cooper, eds. (1977) Dynamic Aspects of Speech Production. (Tokyo: Univ. Tokyo Press).
Verbrugge, Robert and N. S. McCarrell. (1977) Metaphoric comprehension: Studies in reminding and resembling. Cog. Psychol. 9, 494-533.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers SR-21/22 to SR-49:

Status Report		DDC*	ERIC*
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A103325	ED-109-722
SR-42/43	April - September 1975	AD A018369	ED-117-770
SR-44	October - December 1975	AD A023059	ED-119-273
SR-45/46	January - June 1976	AD A026196	ED-123-678
SR-47	July - September 1976	AD A031789	ED-128-870
SR-48	October - December 1976	AD A036735	ED-135-028
SR-49	January - March 1977	AD A041460	ED-141-864
SR-50	April - June 1977	AD A044820	ED-144-138
SR-51/52	July - December 1977	**	**

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

**DDC and/or ERIC order numbers not yet assigned.